International Journal of Computing and Digital Systems

Word sense disambiguation task for Bodo language using Attention based Deep CNN architecture

Subungshri Basumatary[1], Manas Barman[2], Anup Kumar Barman[3], Amitava Nag[4] and Bihung Brahma[5]

[1] *Computer Science and Engineering,Central Institute of Technology Kokrajhar, Kokrajhar, India, 783370*

Abstract: Interest in Natural Language Processing (NLP) has grown very quickly over the last decades, mainly because it provides tools to represent and analyze human languages computationally. A key challenge in NLP is word categorization or classification based on its meaning within a given context. This problem is referred to as word-sense disambiguation (WSD). This issue is prevalent in all languages around the world. However, WSD poses the greatest challenge among North-East Indian languages due to the scarcity of digital resources. This work is an attempt to solve the problem of Word Sense Disambiguation in a low-resource Bodo language and is also considered text-sparse using an adapted Convolutional Neural Network (CNN) model with an attention mechanism. The northeastern region of India predominantly speaks the Bodo language, necessitating careful consideration of its data when constructing NLP models. An attention layer has been implemented in order to effectively identify the significant properties associated with a particular label, enabling the model to focus on the more important things. The CNN layer again extracts certain semantic components from sentences, which further helps in catching subtle nuances of meaning. Testing results were promising, as the proposed framework achieved a remarkable accuracy of 71.43% on a very narrow dataset. Therefore, it demonstrates that the deep CNN with soft attention is more effective in inferring the meaning of words in the Bodo language. Hence, the study proves that NLP, using advanced methodologies like the CNN-Attention model, has immense potential to get over these challenges in low-resource languages. By drawing powerful attention mechanisms and convolutional neural networks together, the model is endowed better at capturing fine-grained semantic differences, offering a glimpse into the possibility for better language processing tools in Bodo and other similarly resource-limited languages.

Keywords: NLP, WSD, Deep learning, CNN, Attention layer and Bodo language.

1. INTRODUCTION

One important sub-branch of artificial intelligence is natural language processing (NLP), concerned with the interactions between computers and human languages. NLP technologies, in their current form, are able to handle quite a diverse array of applications in domains like sentiment analysis, speech recognition, machine translation, and other sophisticated domains. These technologies are core to efficiently processing and examining large amounts of text data. By automating procedures, extracting valuable information from interactions with clients, and increasing the overall efficiency of operations, NLP confers a host of benefits on small-to-medium businesses (SMBs). These include sentiment analysis, which assists in making sense of customers' feelings and feedback; language translation, making any language more accessible to people from other parts of the world; and named entity recognition, for the identification and classification of important information within texts. In addition to this, text classification and text summarization smooth the process of information processing, whereas chatbots and virtual assistants enhance customer service. Keywords will be extracted that aid in identifying pertinent terms for a more thorough examination of the created content; these keywords facilitate text-to-speech and speech-to-text technologies, thereby enhancing communication. Overall, NLP stands at the core of modern AI applications, empowering SMBs to leverage textual data effectively and gain a competitive edge in their respective markets [1].

Despite their linguistic richness, North East Indian languages are still considered low-resource languages due to a lack of digital resources and the scarcity of research conducted in this field. Most NLP technologies, like ChatGPT, primarily develop for widely spoken languages, making it challenging to adapt NLP tools and models to these languages. This disparity highlights the necessity for targeted efforts to bridge the gap. To address this issue, linguists and researchers are currently focusing on creating corpora, developing NLP models, and generating additional language-specific resources

E-mail address: ph19cse1911@cit.ac.in,p23cse1006@cit.ac.in,ak.barman@cit.ac.in,amitava.nag@cit.ac.in,b.brahma@cit.ac.in

tailored to North East Indian languages. These efforts aim to enhance the digital presence and technological accessibility of these languages, enabling the development of NLP tools that cater specifically to their unique linguistic characteristics. The goal is to bring the benefits of advanced NLP applications to speakers of North East Indian languages, thereby preserving and promoting their linguistic heritage in the digital era.

Over the past several decades, there has been significant emphasis in the discipline of computational linguistics and research areas on automated understanding of text. One of the core issues in this field is word-sense disambiguation (WSD), which aims to determine the exact meaning of the word in the context of a given sentence. WSD is an important issue in Natural Language Processing (NLP) since it can have a considerable impact on the efficiency of many language-based applications related to information extraction machine translation, etc. WSD enhances the power and performance of such applications by providing appropriate words for the correct disambiguation of word meanings; hence, it is a very important component of today's advanced NLP systems.

However, WSD is particularly difficult for languages that have a large instructional linguistic variation, such as Bodo. This is one of the common languages spoken primarily in North-East India. The complexity arises due to the many meanings and usage differences of the variations of words in those languages. Unfortunately, this linguistic richness creates complications in the applications of existing NLP tools and models since they are designed to fit more widely spoken and well-resourced languages. Among the very complex languages with respect to syntax and semantics, the Bodo language presents huge challenges; therefore, specific approaches in WSD are absolutely required. According to Mallery [2], WSD is considered an AI-complete problem, similar to NP-completeness in complexity theory [3]. This comparison underscores the considerable computing difficulties associated with achieving effective word sense disambiguation. To illustrate the typical structure of a natural language word sense disambiguation system, Figure 1 presents the general architecture of Natural Language Word Sense Disambiguation (NLWSD). This architecture includes various components designed to analyze context, extract relevant features, and apply algorithms to accurately determine word meanings. Through ongoing research and the development of innovative models, such as attention-based deep CNNs, the goal is to enhance the accuracy and applicability of WSD for linguistically diverse languages like Bodo, ultimately contributing to the advancement of NLP technologies for low-resourced languages.

Machine Learning (ML) and Deep Learning (DL) models have been very popular for resolving a number

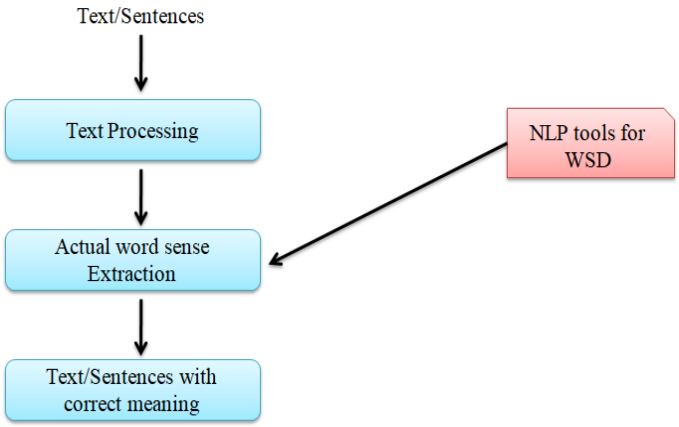

Figure 1. General Architecture of NLWSD.

of problems in the field of NLP, including WSD. Such methods are important for building complex models that have the capacity to efficiently disambiguate word senses with the aid of sizable datasets and contextual information. ML and DL methods have generally improved the accuracy and efficiency of WSD, making them pivotal in modern research on NLP domains. Despite the success of DL techniques for WSD in many languages, the Bodo language still lacks comprehensive exploration of these methods. Thus, this gap highlights the need for focused, specialised research towards adapting such advanced methods for poorly resourced languages like Bodo.

The proposed work mainly focuses on developing a deep learning model exclusively aimed at resolving WSD in Bodo. A lightweight 1D-Convolutional Neural Network (1D-CNN) architecture is presented here for the WSD task. This model aims to classify ambiguous sentences in Bodo accurately. By focusing on a 1D-CNN approach, we can efficiently process sequential data and capture relevant features that contribute to disambiguating word meanings in context. Our study proposes this novel architecture as a step towards improving NLP tools for the Bodo language, thereby addressing the unique challenges posed by its linguistic characteristics and resource constraints. The major contributions of this paper are outlined below:

1) This is the first time a Word Sense Disambiguation model for the Bodo language has ever been attempted. This work contributes to the improvement of NLP resources for low-resource languages and fills a significant need in NLP research for Bodo language.
2) The proposed model introduces an attention-based Deep CNN architecture for WSD. This model leverages DL techniques in order to efficiently distinguish between word senses in the

Bodo language.
3) A sense-tagged dataset tailored to the Bodo language has been created for this work. This dataset not only offers useful information for this study but also paves the way for future research and NLP applications in Bodo.
4) The proposed architecture manages to attain a noteworthy accuracy rate of 71.43% despite the difficulties presented by a limited dataset. This shows how well the attention-based Deep CNN model detects word senses in the Bodo language.

The remaining sections of this article are organised in the following manner. Section 2 gives a general background on previous work in Word Sense Disambiguation. Section 3 outlines the structure of the proposed model, followed by the presentation and examination of results in section 4. Finally, section 5 summarizes the presentation of the study

2. RELATED WORKS

WSD solutions fall into several categories including supervised approach, unsupervised approach, semisupervised approach, knowledge-based approach, DL etc. Methods for knowledge-based WSD rely on lexical databases like WordNet [4]. The Lesk method, which Michael Lesk [5] introduced in 1986, examines the overlap methods between the definitions of the competing terms close to the ambiguous words in that particular text. It is a knowledge-based process that gives an uncertain word its precise meaning using a machine-readable dictionary. Subsequently, Banerjee and Pedersen [6] proposed a modification of Lesk's word sense disambiguation method based on dictionaries. They used lexical database WordNet in place of a traditional dictionary as the source of glosses for their methodology. As a result, this technique achieves an overall accuracy of 32% when tested using English lexical sample data from the Senseval-2 word sense disambiguation exercise. Gautam and Sharma [7] extended Lesk approach on Bigram and Trigram words in Hindi-language sentences that contain verb terms that are polysemic. When it comes to resource-scaring languages, knowledge-based WSD techniques are quite popular [8][9][10]. However, knowledge-based WSD approaches suffer from the overlap sparsity and dimensional explosion problem [11].

In supervised WSD approaches, the WSD classifier is trained using labelled data. On the other hand, Unsupervised WSD techniques cluster unlabeled corpus to identify the semantic category of the unclear word. Recently, many works on WSD have used both supervised [12][13] and Unsupervised [14][15] WSD techniques. In certain articles [16][17], semi-supervised WSD methods that blend labelled and unlabeled data for learning were proposed. The dimensional explosion issue in conventional NLP models was first addressed by Bengio et al. [18], who employed neural networks to train models for vast quantities of text and map textual words to N-dimensional vector spaces. A novel neural network architecture that embeds multi-relational graphs into a flexible continuous vector space while maintaining and improving the original data [19]. Collobert et al. [20]introduced CNN into the field of NLP for the first time. Subsequently, Kim [21] enhanced the performance of the CNN for text classification.

Although deep neural network (DNN) techniques for word sense disambiguation (WSD) might produce better results, current deep learning (DL) strategies predominantly focus on other applications within the natural language processing (NLP) domain. These applications include tasks such as rumor detection [21], question categorization [22], and text classification [23]. While a few DL-based studies have attempted to address WSD for various resource-rich languages [24], there remains a significant gap in research regarding effective solutions for WSD in the Bodo language. To the best of our knowledge, no successful methods have been proposed for WSD in Bodo, whether using DL or other techniques.

To address this gap, we propose a lightweight one-dimensional convolutional neural network (1D-CNN) architecture tailored for the WSD task in the resource-limited Bodo language. Our proposed approach aims to provide a feasible and efficient solution for disambiguating word senses in Bodo, leveraging the strengths of CNNs in processing sequential data. The 1D-CNN's lightweight nature makes it particularly suitable for scenarios with limited computational resources, which is often the case for under-resourced languages like Bodo. By focusing on this specific challenge, we aim to contribute to the broader field of NLP and improve the accessibility and accuracy of language processing tools for Bodo, paving the way for further research and development in this area.

3. PROPOSED METHODOLOGY

Figure 2 illustrates the schematic representation of proposed framework for WSD task. A collection of datasets is generated by human means for the purpose of data preparation. A CNN model is created following the implementation of pre-processing techniques such as punctuation removal, tokenization, and removal of stop word. The pre-processing of word embeddings is performed using the Keras library. The methodology employed in this paper encompasses several key steps:

A. Sense-Tagged Dataset Preparation

In this work, we have developed a dataset of ambiguous words in the Bodo language, annotating each word with its respective senses based on contextual information. For example, consider the statement **"गितानि खानायआ गोजा गाबनि"** (Geeta has red hair color). In this statement, the word **"खानाय"** is ambiguous, as it has two distinct meanings: one referring to "hair" and the other

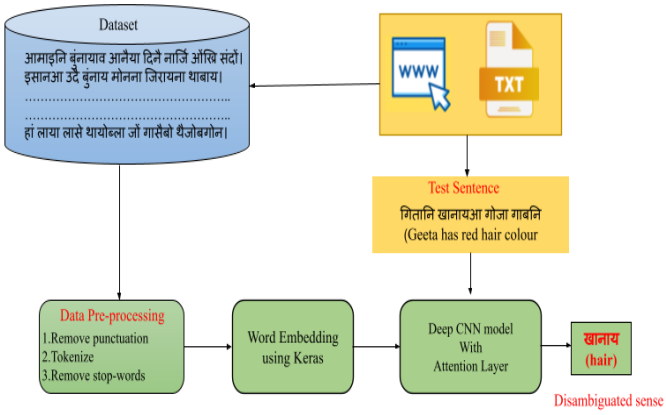

Figure 2. Schematic representation of proposed framework for WSD task.

denoting "to tide." To address such ambiguities, we have created a labelled dataset comprising multiple sentences that contain ambiguous words. The dataset annotates each ambiguous word with its potential meanings, facilitating enhanced comprehension and context-based disambiguation. This dataset is valuable for natural language processing tasks, such as machine translation and text analysis, as it helps in accurately interpreting the meaning of ambiguous words in different contexts. The dataset aims to improve the precision and effectiveness of Bodo language processing tools. For instances a table has been created with different classes which is depicted in Table I.

B. Data pre-processing

The sentences within the database exhibit a high degree of diversity and undergo preprocessing prior to the use of CNN method. Preprocessing architecture necessitates the implementation of three distinct phases.

1) Punctuation removal: To build the databases, we generate words and test them using our testing data. Consequently, the punctuation marks including (',', ' ', '(',')', '-', '.....',:', "'",) are removed from our dataset.
2) Tokenize: In order to facilitate the processing of the input, sentence tokenization is employed to break down phrases into individual words. For instances as- **"गितानि खानायआ गोजा गाबनि"** (Geeta has red hair color) after tokenization output will be: **"गितानि" "खानायआ" "गोजा" "गाबनि".**
3) Stop word removal: The process of stop word reduction involves eliminating words that are frequently used in sentences but do not contribute any significant meaning when used to construct a database. For instance, the words **आरो** (and), **एबा** (or), **जेरैनों** (who's), and **सोर** (who) do not provide any substantial meaning to the sentences without proper contextualization.

C. Word Embedding

Word embeddings are an essential technique in Natural Language Processing (NLP) that allows words to be represented as compact vectors in a continuous vector space. These embeddings precisely capture the semantic associations between words, rendering them extremely efficient for a range of NLP tasks, including text classification. Keras, a widely used deep learning framework, provides straightforward utilities for incorporating word embeddings into models designed for text classification. One such utility is the embedding layer, which transforms words (or tokens) represented as numeric values into compact vectors of a predetermined dimension. This transformation allows the model to learn and leverage the semantic information embedded in the word vectors, thereby improving the performance of text classification tasks. By using word embeddings, the model can better understand and process the contextual meanings of words, leading to more accurate and efficient NLP applications.

D. Attention-Based Deep CNN Model

The model can successfully pick up and use contextual signals thanks to the design and implementation of a Deep Convolutional Neural Network (CNN) architecture that includes attention mechanisms. This combination results in a powerful deep learning model tailored for text processing tasks. A CNN with an attention mechanism leverages the strengths of both components to enhance its performance. The CNN's feature extraction capabilities enable it to identify and analyse local patterns within the text, such as specific phrases or word combinations. Meanwhile, the attention mechanism focuses on the most relevant parts of the text, allowing the model to consider the global context and prioritise important information. By combining these elements, the model can effectively extract both local and global context from text data, leading to a more accurate and nuanced understanding. This approach is particularly beneficial for complex NLP tasks, such as sentiment analysis, text classification, and word sense disambiguation, where understanding the broader context is crucial. Adding attention mechanisms to CNNs is a big step forward in the field of natural language processing (NLP) because it creates a strong framework for processing and understanding text data more accurately and in more depth. Figure 3 illustrate the proposed attention-based deep Convolutional neural network (CNN) model.

1) Convolution and pooling layer

DL neural networks are increasingly used in text classification recently in order to extract semantic information from large amounts of text input. A general feedforward neural network can be represented as follows:

Table I. Instances of sense tagged data

Sl No.	Sentences(Q_L)	Classes (S_L)
1	गितानि **खानायआ** गोजा गाबनि (Geeta has red hair colour)	खानाय (hair)
2	राहुलआ मोसौ **खानायखौ** हगारना होयो (Rahul released the tied cow)	खानाय (to tide)
3	समबारूआ एमबु **बारनाय** गेलेदों (Sambaru is playing frog jump)	बार (jump)
4	आसामाव बोसोरफोरोमबो बारहुंखा **बार** बारो (Every year there is a storm in Assam)	बार (wind)
5	राहुलनि लोगोनि **बिसिआ** आमेरिकानि (Rahul's friend's wife is from American)	बिसि (wife)
6	मैनाया गोजाम बिजाबफोरखौ **बिसिना** गाहर जोबबाय (Mwina tear off and threw away all her old books)	बिसि (tear off)

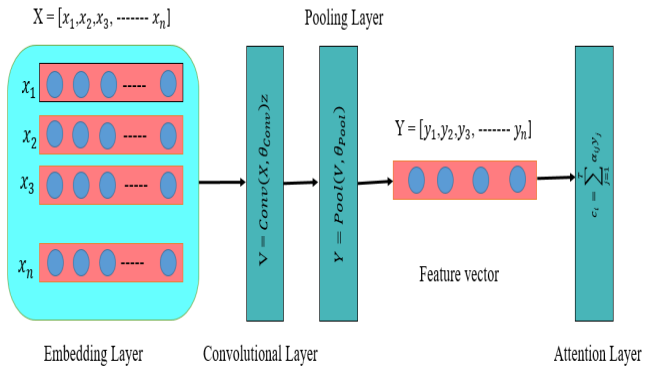

Figure 3. Proposed Attention base CNN architecture for WSD task.

$$Y = F(X, \theta) \tag{1}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ represent a set of inputs, $Y = \{y_1, y_2, \ldots, y_m\}$ is a set of outputs. F and θ represent a feedforward neural network model and the set of parameters in the model.

Convolutional neural networks (CNNs) are a well-known variant of feedforward neural networks (FNNs) that utilise a deep architecture with convolutional operations. They have proven highly effective in the domain of natural language processing (NLP), specifically for tasks related to text mining. CNNs thrive on recognizing localized and position-invariant patterns in text, making them particularly well-suited for tasks that involve identifying unique phrases or word combinations, regardless of their position within the text. As a result, CNNs have become a popular deep learning model for text mining applications.

The convolution layer, the fundamental building block of convolutional neural networks (CNNs), consists of several convolution kernels. The kernels execute the convolution computation by smoothly moving across the local windows of input data, extracting relevant characteristics at each iteration. This technique enables the model to efficiently capture and analyse local dependencies and patterns present in the text. The pooling layer, which comes after the convolution layer, aims to eliminate a sample from the convolutional output, reduce the dimensionality of the convolution vectors, and prevent overfitting. By summarising the characteristics retrieved by the convolution layers, pooling enhances the model's robustness and performance by providing a more comprehensive representation of the data. Overall, the combination of convolution and pooling layers enables CNNs to process and analyse textual data efficiently, making them a powerful tool in NLP. In light of this, Y=F(X, θ) of CNN with convolution and pooling layer can be represented as follows:

$$Y = \text{Pool}(\text{Conv}(X, \theta_{\text{Conv}}), \theta_{\text{Pool}}) \tag{2}$$

where Conv stands for the convolutional layer and θ_{Conv} for the convolutional layer's set of parameters. Pool denotes the pooling layer, and θ_{Pool} denotes the pooling layer's set of parameters.

Pooling can be categorised into two distinct types: maximal pooling and average pooling. Maximum pooling is highly effective in text categorization because it saves crucial textual information, while average pooling maintains the average values of features in the feature map. Our proposed model incorporates maximum pooling to preserve the most significant characteristics, thereby improving the model's performance in text

analysis.

2) Attention mechanism layer

An attention layer is an essential component in NLP activities like text mining and text categorization. This layer's inclusion is particularly beneficial for tasks involving the processing of text sequences of varying lengths. It enables models to concentrate on certain segments of the input text while making predictions. By assigning weights to different features, the attention layer could help in highlighting various sections of the text that are important to the model's prediction. This approach improves the ability of the model to extract relevant information and improves performance. This paper embeds the attention block in CNN's network, particularly after a pooling layer. This placement ensures that the model extracts not only essential features through convolution and pooling but also pays attention to the most crucial features before predicting. This model can now combine these complex and heterogeneous inputs, such as text, to have a better representation of text factors for higher accuracy and interpretability of tasks related to text mining and categorization. An attention block that has the potential to learn weights automatically, as is shown by its output in Equation 3.

$$c_i = \sum_{j=1}^{T} \alpha_{ij} y_j \tag{3}$$

c_i and y_j denotes the global features and the output of its attention block respectively.

4. Experimental Results and Analysis

In this section, we will attempt to discuss the outcome of the result through the following steps. In the initial phase, this section reviews all the results produced by a complex classification model trained with a very long chain of training epochs. The primary objectives of this model are to put forward tasks for the correct determination and classification of text inputs into the various classes of words in Bodo. A thorough analysis is required to evaluate the model's performance and understand its ability to masterfully handle its complexities during Bodo language classification. By analyzing the results, we can assess the model's effectiveness in identifying and categorizing various Bodo word classes, constituting: 'बुंनाय', 'खानाय', 'हाबा', 'बिसि', 'गं', 'जा', "आ'दै", 'हा', 'बार', 'हर', 'गाब', 'अर', 'समाय', 'गुदुं', 'गाव', 'खन', 'फुं', 'लाउ', 'खुसेर', 'जि',' एव', 'हाग्रा',' 'हांखो', 'जराय', 'गोमो', 'हां', 'खाम', 'मोसा', 'महर', 'मान'. with two semantic categories 30 ambiguous words are selected.

A. Evaluation Measure

One of the most commonly used metrics to evaluate performance in WSD tasks is F1 and accuracy. Both metrics convey full knowledge of the model's effectiveness in performing well in properly identifying and classifying word senses, hence managing robust performance in WSD applications [25].

1) Accuracy (A): Accuracy measures the quantitative value of how much the estimated value agrees with its correct or observed value for a model. It is the proportion of correctly classified samples. The accuracy of the model can be determined using the formula as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where, TP = (total positives) , TN = (total negatives) FP = (false positives) and FN = (false negatives).

2) Precision (P): Precision is the ratio of the number of correctly identified positive cases against all examples depicted as positive; it measures the accuracy of the positive predictions.

$$P = \frac{TP}{TP + FP} \tag{5}$$

where, P = precision, TP = (total positives) and FP = (false positives).

3) Recall (R): Recall is a measure that expresses the ratio of correctly detected cases of the true positive class out of all actual positive examples; it therefore measures the completeness of the retrieval of positive occurrences.

$$R = \frac{TP}{TP + FN} \tag{6}$$

where, R = recall, TP = (total positives) and FN = (false negatives).

4) F1-score: The F1 score is frequently interpreted as the weighted average of precision and recall multiplied by the harmonic mean of accuracy and memory. The following formula is used to determine the F1-score:

$$F1\text{-}score = \frac{2 \cdot (P \cdot R)}{P + R} \tag{7}$$

where, P = precision and R = recall.

Table II presents some examples of Bodo words that exhibit ambiguity, with similar meanings, and the classification report showing the details of accuracy calculation, precision, and recall. Through the analysis of these examples and their corresponding classification evaluations, one cannot truly realise the various nuances and pitfalls likely to occur while interpreting the ambiguities of the Bodo language. This analysis brought out the effectiveness of these word differentiations that could help in the use and making of better language processing and translation tools for Bodo and more

accurate, contextually correct communication.

Here Figure 4 and 5 demonstrate the calculation of graphs for both training and validation accuracy, as well as the corresponding loss for training and validation. These graphs provide a clear visual representation of the model's performance and its learning progress over the training period.

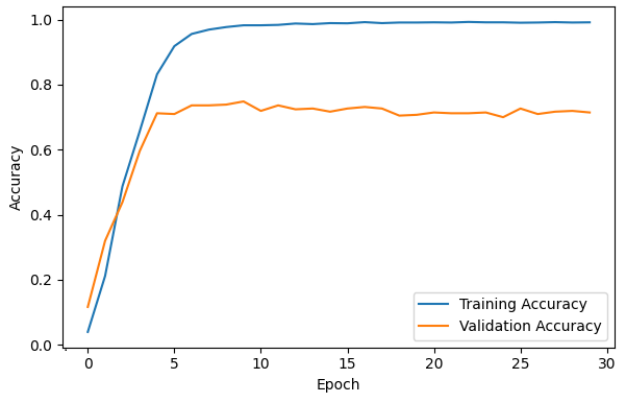

Figure 4. Evaluation graph for Training and Validation Accuracy.

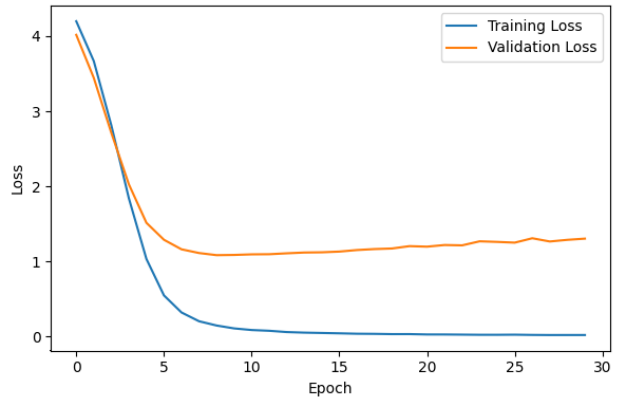

Figure 5. Evaluation graph for Training and Validation Loss.

B. Comparative analysis

In this section, we present a detailed contrast between our proposed work and recent studies on the word-sense disambiguation task. We trained a proposed CNN-based attention deep learning model using a dataset that contains 30 ambiguous words, then recorded the validation accuracy and F1 measure as shown in Table III. In contrast, two other works make use of specifically compiled datasets of 66 and 50 ambiguous words in the Panjabi and Urdu languages, respectively. We will also evaluate the performance using two common WSD metrics: F1 and accuracy. Both give a full picture of how good the model is at correctly identifying and grouping word senses, resulting in strong and reliable performance in WSD applications. Table III reveals that our study's dataset includes 30 ambiguous words, each with an average of 27.8 instances and 2.4 senses. This guarantees a balanced and diverse dataset for effective WSD. Despite using a much smaller dataset compared to others, our framework posted the highest F1-score against all systems, which proves the strength and accuracy of the handling methodology for ambiguity in text.

Figure 4 and 5 illustrate both training and validation evaluation graphs with accuracy and loss metrics. High validation accuracy and F1-scores will prove that our CNN-based attention model captures the fine distinctions of ambiguous words effectively and generalizes very well to unseen data. This instantiation makes it salient that our framework is robust enough to accomplish higher performance in WSD tasks even with a limited dataset and creates a benchmark for future studies in this area.

This comparative study highlights our method as being more effective in Word Sense Disambiguation (WSD), especially when limited data is used. Our model demonstrated a superior focus on F1-scores for 30 ambiguous words in the limited-resource Bodo language, demonstrating its potential and robustness in WSD. It has thus shown the capacity for efficiently dealing with ambiguity even when limited data is available, making it very promising for other low-resource languages. These results demonstrate the fitness of our CNN-based Attention Model in accurately capturing complex word meanings in contexts, setting a very strong precedent for future WSD research.

5. Conclusion

In the realm of natural language processing (NLP), this study marks a crucial turning point, particularly for low-resource languages such as Bodo. It proposes an attention-based deep CNN model for word sense disambiguation in Bodo and fills a prominent gap in existing research. The construction of the sense-tagged dataset and the development of this world-first model laid a very strong foundation underpinning improved NLP applications and far deeper language understanding for Bodo. Structurally, the proposed architecture achieves an accuracy rate of 71.43% despite constraints resulting from data limitations. This promising result highlights the model's potential for various NLP applications and makes a substantial contribution to the conservation and understanding of the Bodo language. By leveraging advanced techniques like attention mechanisms and convolutional neural networks, this study paves the way for future research and development in NLP for other similarly resource-limited languages.

Table II. Evaluation of the classification report of some ambiguous words

Ambiguous word	Precision	Recall	F1-score
बुंनाय	1.00	0.89	0.94
हां	0.93	0.93	0.93
हाबा	0.73	0.89	0.80
बिसि	0.94	0.73	0.82
गं	0.96	0.96	0.96
जा	0.95	1.00	0.98
आ'दै	0.86	0.95	0.90
हा	0.88	0.79	0.83
बार	0.91	0.91	0.91
हर	0.96	0.96	0.96

Table III. Comparative analysis of the proposed work with related works

References	Languages	Dataset	No. of Ambiguous word	Instance per sense	Average sense per word	Model	Accuracy (%)	F1-score
[26]	Punjabi	custom	66	25.7	4.2	LSTM	71.5	Not given
[25]	Urdu	custom	50	Not given	Not given	LSTM	72.63	60.0
Proposed work	Bodo	custom	30	27.8	2.4	CNN Attention Layer	71.43	67.11

References

[1] F. Alam, A. Hasan, T. Alam, A. Khan, J. Tajrin, N. Khan, and S. A. Chowdhury, "A review of bangla natural language processing tasks and the utility of transformer models," *arXiv preprint arXiv:2107.03844*, 2021.

[2] J. C. Mallery, "Thinking about foreign policy: Finding an appropriate role for," 1994.

[3] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.

[4] L. Huang, C. Sun, X. Qiu, and X. Huang, "Glossbert: Bert for word sense disambiguation with gloss knowledge," *arXiv preprint arXiv:1908.07245*, 2019.

[5] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proceedings of the 5th annual international conference on Systems documentation*, 1986, pp. 24–26.

[6] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in *International conference on intelligent text processing and computational linguistics*. Springer, 2002, pp. 136–145.

[7] C. B. S. Gautam and D. K. Sharma, "Hindi word sense disambiguation using lesk approach on bigram and trigram words," in *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, 2016, pp. 1–5.

[8] A. Haque and M. M. Hoque, "Bangla word sense disambiguation system using dictionary based approach," *ICAICT, Bangladesh*, 2016.

[9] A. R. Pal and D. Saha, "Word sense disambiguation in bengali: an unsupervised approach," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2017, pp. 1–5.

[10] D. Das Dawn, A. Khan, S. H. Shaikh, and R. K. Pal, "A dataset for evaluating bengali word sense disambiguation techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 4057–4086, 2023.

[11] H. Hwangbo and Y. Kim, "An empirical study on the effect of data sparsity and data overlap on cross domain collaborative filtering performance," *Expert Systems with Applications*, vol. 89, pp. 254–265, 2017.

[12] S. Yamaki, H. Shinnou, K. Komiya, and M. Sasaki, "Supervised word sense disambiguation with sentences similarities from context word embeddings," in *Proceedings of the 30th Pacific Asia conference on language, information and computation: oral papers*, 2016, pp. 115–121.

[13] A. Saif, N. Omar, U. Z. Zainodin, and M. J. Ab Aziz, "Building sense tagged corpus using wikipedia for supervised word sense disambiguation," *Procedia Computer Science*, vol. 123, pp. 403–412, 2018.

[14] A. R. Pal and D. Saha, "Word sense disambiguation in ben-

gali language using unsupervised methodology with modifications," *Sādhanā*, vol. 44, pp. 1–13, 2019.

[15] B. Moradi, E. Ansari, and Z. Žabokrtskỳ, "Unsupervised word sense disambiguation using word embeddings," in *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, 2019, pp. 228–233.

[16] P. Rani, V. Pudi, and D. M. Sharma, "Semisupervied data driven word sense disambiguation for resource-poor languages," in *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 2017, pp. 503–512.

[17] ——, "Semisupervied data driven word sense disambiguation for resource-poor languages," in *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 2017, pp. 503–512.

[18] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[19] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "A semantic matching energy function for learning with multi-relational data: Application to word-sense disambiguation," *Machine Learning*, vol. 94, pp. 233–259, 2014.

[20] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[21] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[22] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, and A. Khattak, "Exploring deep neural networks for rumor detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4315–4333, 2021.

[23] J. Liu, Y. Yang, S. Lv, J. Wang, and H. Chen, "Attention-based bigru-cnn for chinese question classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2019.

[24] A. Kenarang, M. Farahani, and M. Manthouri, "Bigru attention capsule neural network for persian text classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 8, pp. 3923–3933, 2022.

[25] A. Saeed, R. M. A. Nawab, and M. Stevenson, "Investigating the feasibility of deep learning methods for urdu word sense disambiguation," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 2, pp. 1–16, 2021.

[26] V. P. Singh and P. Kumar, "Word sense disambiguation for punjabi language using deep learning techniques," *Neural Computing and Applications*, vol. 32, pp. 2963–2973, 2020.

SUBUNGSHRI BASUMATARY , completed her B.Tech degree from Central Institute of Technology Kokrajhar from the Department of Computer Science and Engineering. Also received her M.Tech degree from Tezpur University from the Department of Information Technology. Now pursing her PHD degree from Central Institute of Technology Kokrajhar. Her research interests include Machine Learning, Deep Learning Natural Languages Processing(NLP) and Word Sense Disambiguation (WSD) their applications in the low resources languages.

MANAS BARMAN , was born in Golakganj,Dhubri,Assam. He completed his B.Tech degree in Computer Science and Engineering from the Central Institute of Technology Kokrajhar, Kokrajhar, Assam, in 2023. Currently, he is pursuing his M.Tech degree in Computer Science and Engineering from the same institution. His research interests include machine learning, deep learning, natural language processing as well as IoT and cybersecurity.

ANUP KUMAR BARMAN (Member, IEEE) received the M.Sc., M.Tech., and Ph.D. degrees from Gauhati University, Assam. He is currently an Assistant Professor of computer science and engineering with the Central Institute of Technology Kokrajhar, Assam, India. He is actively engaged in a project called ''CLIA-Cross-Lingual Information Access'' executed in collaboration with various reputed institutes, such as Gauhati University, IIT Mumbai, IIT Hyderabad, and IIT Kharagpur. He is also engaged in various research activities, such as the development of the stemmer, word sense disambiguation module and parser, and so on for the Assamese language. He has more than 20 research publications in various international journals and conference proceedings.His research interests include natural language processing, machine learning, information retrieval, and information security.

AMITAVA NAG (Senior Member, IEEE) is currently a Professor of computer science and engineering with the Central Institute of Technology Kokrajhar, Assam, India. He has more than 70 research publications in various international journals and conference proceedings. His research interests include the IoT, information security, and machine learning. He is a fellow of IEI.

BIHUNG BRAHMA is an Assistant Professor in the Department of Humanities and Social Sciences at Central Institute of Technology Kokrajhar. His studies and research area are language and linguistics. He is presently associated with Bodo language and cultural development such as 'Bodo and Dimasa Heritage Digital Archive' and contributing as a language expert, annotator and reviewer for Bodo in AI4Bharat project initiated by IIT Madras.