ISSN 2007-9737

Named Entity Recognition (NER) for Sino-Tibetan Languages: A Comprehensive Review and Status

Jinia Angeline Gayary*, Shikhar Kumar Sarma, Hiren Kumar Deva Sarma, Kuwali Talukdar

Gauhati University, Department of Information Technology, Guwahati,
India

jingayary51@gmail.com, sks001@gmail.com, hirenkdsarma@gmail.com, kuwalitalukdar@gmail.com

Abstract. As technology continues to advance at a rapid pace, there is a growing interest in Natural Language Processing (NLP) tools and applications. However, creating NLP tools that can effectively process natural languages presents numerous difficulties. One crucial aspect of NLP is Named Entity Recognition (NER), which involves identifying and classifying named entities in a text based on their surrounding context. Although there has been extensive research, NER tagging still struggles to accurately tag unfamiliar named entities. NER for Sino-Tibetan languages, such as Bodo and Myanmar, poses various challenges, including word segmentation, lack of resources, and ambiguity. In this paper, we review the state-of-the-art in NER for Sino-Tibetan Languages, focusing on the methods, datasets, and performances achieved. We also highlight underlying issues and future directions for NER research in this domain. Although there are not many works on NER related to Sino-Tibetan languages available, we tried to cover a good number of papers with a wide spectrum of languages, so that this review could be best utilised by researchers interested in NER studies and development for language technologies for languages from this group. As many as different works on Sino-Tibetan NER studies have been covered. We also tried to cover NER works with a variety of approaches and techniques ranging from rule-based to machine learning, deep learning, hybrid, and cross-lingual methods and highlighting their relevance towards the specific linguistic demands of Sino-Tibetan languages. Apart from these, we also reviewed a brief status on the NLP tasks for low-resourced languages, Bodo and Assamese. We have analyzed and presented in a structured way all the approaches, methods used, along with datasets, performances and challenges encountered. We hope that this paper can provide a comprehensive overview and a useful resource for the research community interested in NER for Sino-Tibetan languages.

Keywords. Named entity recognition, natural language processing, Sino-Tibetan language, deep learning, low-resource language.

1 Introduction

Artificial Intelligence (AI) is greatly influencing modern life within today's internet culture. One such field in AI where progress has gradually increased as needed is in Natural Language Processing (NLP), as it helps process and generate textual data to a large measure. NLP deals with the analysis and generation of natural language. It enables us to use various applications in different domains that can enhance human communication, information access, and knowledge discovery.

However, NLP still faces many challenges, such as the diversity and complexity of natural languages, data scarcity and quality, and ethical implications of language technologies. One of the core tasks of NLP is Named Entity Recognition (NER), which is the process of identifying and classifying named entities from an unstructured text. NER can be beneficial for various applications, such as information extraction, question answering, machine translation, and sentiment analysis. NER also plays a vital role in content moderation, where tasks like hate speech detection and toxic comment identification rely on accurate entity recognition as seen in Assamese [4, 39]. Text summarization [23] benefits from entity-aware processing that helps retain key information [16, 7, 5]. The development

doi: 10.13053/CyS-29-3-5896

of supporting resources like Assamese and Bodo wordnets [52, 51, 6], lemmatizers [15] and stemmers [53] has significantly contributed to semantic processing and NER. Tools such as syllable-based word segmentation [46] and back-transliteration systems [35] aid part-of-speech tagging, enabling better adaptation of NER systems to low-resource scenarios. Techniques like feature selection [40] and multiword expression identification [45, 42, 41] further support precise boundary detection for complex entities. The complex linguistic characteristics of the Sino-Tibetan language family make NER for these languages especially difficult.

Sino-Tibetan languages are a large and diverse family of languages spoken by around 1.4 billion people in Asia, mainly in China, India, and Southeast Asia. Sino-Tibetan languages include Chinese, Tibetan, and other related languages, such as Bodo, Manipuri, and Myanmar etc. These languages have various linguistic features and challenges that make NER a difficult task, such as word segmentation, orthographic variation, tonal variation, and multilingualism. Among the Sino-Tibetan languages, Chinese is the most widely spoken and studied language, having received the most attention and research in NER. There are many methods, datasets, tools, and evaluation metrics for Chinese NER, which have achieved high accuracy and state-of-the-art results. These include both general-purpose NER systems and domain-specific efforts, such as in medical [27, 2, 58] and cybersecurity [17] domains, which, though not directly aligned with general-purpose NER, introduce valuable architectures like BiLSTM-CRF, attention mechanisms, and dictionary-based techniques that offer transferable insights for low-resource languages like Bodo. However, Bodo, a key focus of this review, is the least resourced Sino-Tibetan languages. It is spoken widely in Assam, India and parts of neighbouring nations like Nepal and Bangladesh. Given its unique linguistic and cultural features, Bodo remains unexplored to a great extent in foundational NLP tasks. Assamese, an official language of Assam, shares historical and linguistic connections with Bodo and plays a supportive role in resource creation for low-resource scenarios. Studies in Assamese-Bodo Neural Machine Translation (NMT) [56] emphasize the need for high-quality bilingual corpora and resource modelling [1], both of which are important for training robust NER systems in low-resource settings. Corpus-building efforts for Bodo [10] and Manipuri [34] highlight valuable linguistic resources that can directly support NER tasks. Foundational tools like Part-of-speech (PoS) tagging and Word Sense Disambiguation (WSD) play a critical role in enhancing NER accuracy. PoS tagging has been explored for Bodo [8], Manipuri [43] and Assamese [55], while WSD approaches using genetic algorithms and statistical methods have shown potential for Assamese. These efforts yield grammatical cues and semantic context crucial for languages with complex morphological and ambiguous structures, such as many in the Sino-Tibetan family. Recent work by Liu et. al [30] offers an extensive overview of NLP work done in Sino-Tibetan low-resource languages, identifying the chronic under-representation of these languages on key NLP tasks. Though their work presents a general landscape view, our attention is concentrated on Named Entity Recognition (NER) in particular, which is a fundamental NLP task not yet well-explored for this language family.

With increasing needs for effective techniques of natural language processing, specifically for Sino-Tibetan languages, such as those that are lesser resourced, like Manipuri, Tibetan, Mizo, Bodo, Myanmar, and Kokborok, the need to review Named Entity Recognition for these languages has come up. It is, therefore, of interest in how to deal with these linguistic and resource-poverty unique challenges. Most Sino-Tibetan languages do not have enough labeled datasets to train good NER models. Many advancing areas remain underexplored, like medical NER, network and security domains, which are still at a very nascent stage; hence, this sector requires individually targeted research. This paper presents the very first extensive review of NER methods that specifically apply to the Sino-Tibetan languages.

In contrast to existing reviews on NER, this paper is unique because prior works tend to concentrate most heavily on the well-examined English or

Chinese; yet, in consolidating research over a massive set of Sino-Tibetan languages, which are less-resourced. This work categorizes and analyzes approaches that cut across rule-based, machine learning, deep learning, and cross-lingual methodologies, not only categorizing methods that work but also uncovering gaps within current techniques, datasets, and tools that are particular to Sino-Tibetan languages.

The rest of this paper is organized in the following way: The Search Methodology outlines the approach used to identify and select relevant literature for this review. The Selected NER Works on Sino-Tibetan Languages section gives an overview of the existing research in NER for Sino-Tibetan languages, concentrating on deep learning architectures such as the Bi-LSTM-CRF model. The Performance Evaluation Criteria for Named Entity Recognition (NER) section discusses standard evaluation metrics and aggregation strategies, highlighting considerations specific to Sino-Tibetan languages. The Analysis section offers a comparative summary of different methods, with languages, datasets, and their respective performance scores, along with a graphical analysis at the end to analyze the F-scores for different NER approaches for Sino-Tibetan languages. The Issues and Challenges of NER in Sino-Tibetan languages section presents the challenges faced in developing NER for Sino-Tibetan languages. Finally, the Conclusion and Future Work highlights the most vital findings and provides directions for future research on NER for Sino-Tibetan languages.

2 Search Methodology

To identify and compile studies relevant for this review, we performed a systematic literature search on academic platforms like Google Scholar, Semantic Scholar, IEEE Xplore, Scopus, and ACL. The search spanned the year range from 2010 to 2025 and used keyword searches with combinations of ”Named Entity Recognition,” ”NER in Sino-Tibetan languages,” ”deep learning for NER,” ”NER in low-resource languages,” and language-specific keywords like ”NER in Bodo,” ”NER in Manipuri,” ”NER in Tibetan”. Although Sino-Tibetan languages were our main area of interest, we also included closely related low-resource languages because of their applicability in multilingual and cross-lingual NER studies.

These languages tend to depend on core NLP building blocks such as part-of-speech tagging, transliteration, and lexical resources that are essential in NER system construction. Altogether, 70 papers were examined, and the results are summarized in two tables of summaries: Table 1 provides a breakdown of NER research into Sino-Tibetan and low-resource languages, whereas Table 2 categorizes representative NER works by methodological approach.

3 Selected NER Works on Sino-Tibetan Languages

NER is a key task in natural language processing (NLP), and, in many ways, its development varies across language families. For Sino-Tibetan languages, where annotated corpora are limited and complex linguistic structures are high, selecting an appropriate method becomes critical. While Chinese NER has advanced significantly due to abundant data and tools, languages like Tibetan, Manipuri, Mizo, Bodo, Kokborok, and Myanmar remain underexplored. Table 3 shows the merits and demerits of different NER approaches. These approaches have evolved from rule-based methods to complex deep learning architectures, with each offering distinct advantages depending on resource availability and language characteristics.

The sections below outline these main approaches.

3.1 Rule-Based Approaches

Rule-based methods were the first approaches used for Named Entity Recognition (NER) in resource-scarce settings. Such systems rely on handcrafted linguistic rules and heuristics to identify as well as categorize entities from natural language. Typically, such systems include morphological patterns, part-of-speech (PoS) tags, gazetteers, suffixes, and syntactic cues to build models that match specific entity forms. Their

Table 1. List of Explored Sino-Tibetan Languages and Corresponding Number of NER Publications

Sl. No.	Language	No. of Papers (Year-wise)	Remarks
1	Chinese	12 (2017–2024)	Extensive research using BiLSTM-CRF, attention mechanisms, and transformer-based models; includes general-purpose NER as well as domain-specific applications in medical and cybersecurity.
2	Manipuri	7 (2010–2023)	Sustained research using CRF, BiLSTM, and hybrid methods; strong representation despite being a low-resource language.
3	Bodo	2 (2024–2025)	Emerging research interest with a recent paper using LSTM; remains largely underexplored.
4	Myanmar	4 (2017–2020)	Utilizes syllable- and character-based models with CRF, HMM, and CNN-BiLSTM-CRF tailored to syllabic structure.
5	Tibetan	3 (2010–2019)	Limited studies using CRF, Maximum Entropy, and rule-based methods; still under-resourced for NER.
6	Kokborok	1 (2015)	Initial work using rule-based and MIRA-supervised learning; represents early exploration.
7	Mising	1 (2016)	SVM-based approach with a small annotated corpus; limited follow-up research.
8	Mizo	1 (2016)	Rule-based linguistic recognition attempted; no substantial expansion observed.

form tends to utilize hand-crafted templates and regular expressions that are linguistically tailored to aspects such as honorifics, common suffixes, or grammar forms.

Rule-based methods are especially attractive in resource-poor contexts like those for the Sino-Tibetan language family, where limited annotated data is available for machine learning systems.

Their interpretability, language-specific customizability, and minimal data requirements make them suitable for initial development phases. For example, a suffix-rule-based system attained an F1-score of 83.18% for Kokborok with a straightforward dictionary-based technique [48].

Parallelly, rule templates based on case-auxiliary grammar were used for Tibetan NER [60], and rule-based systems for Mizo were designed from news corpora based on entity-specific patterns [9].

These strategies demonstrate that with a good understanding of the linguistic structure, decent performance can be attained with few resources.

Yet, rule-based systems are apt to fail in coping with ambiguity, irregularities, and exceptions in syntax and often meet scalability issues when extended to diverse domains or dialects.

In addition, rule sets are less practical for large or dynamic applications because maintaining and enlarging them is time-consuming, requiring deep linguistic expertise.

Table 2. Representative NER Works based on different approaches for Sino-Tibetan Languages

Approach	Representative Works	Remarks
Rule-based	[48] (2015), [9] (2016), [60] (2010)	Rule-based systems offer high precision by using linguistic rules, gazetteers, and dictionaries. However, they struggle with ambiguity, lack scalability, and are difficult to maintain. Their rigidity makes them less adaptable to unseen data or evolving language usage.
Machine Learning-based	[44] (2011), [54] (2010), [19] (2016), [29] (2018), [25] (2019), [32] (2017)	Machine learning methods like CRF and SVM have proven effective for NER in low-resource, morphologically rich languages. Their success relies heavily on well-designed linguistic features such as affixes, context, and PoS tags. CRF excels at sequence labeling, while SVM performs well in complex entity types like multiword and reduplication. Careful feature engineering and language-specific adaptation are key to achieving strong NER performance.
Deep Learning-based	[37] (2024), [28] (2023), [62] (2019), [47] (2017), [33] (2020), [24] (2020)	Deep learning models like BiLSTM-CRF, CNN-LSTM, and GRU-CRF capture contextual and sequential patterns using word- and character-level embeddings. Attention and GCNs improve recognition of multiword and nested entities. These models perform well for agglutinative and low-resource languages but struggle with transliterated or rare entities and require significant computational resources.
Hybrid-based	[21] (2023), [57] (2023), [18] (2022), [2] (2022), [61] (2020), [17] (2021), [58] (2019), [22] (2013), [59](2019), [20] (2016)	Hybrid approaches combine linguistic features with deep learning models to boost accuracy. They handle agglutinative languages and OOV issues well using character-level embeddings, but struggle with multiword and transliterated entities. Overall, they balance structure and context effectively.
Cross-lingual	[49] (2020), [31] (2019), [38] (2025)	Cross-lingual methods enable zero/few-shot NER in low-resource Sino-Tibetan languages by transferring knowledge from high-resource languages using shared embeddings. Their success depends on translation quality, alignment accuracy, and linguistic similarity. Prompt-based approaches offer better tagging performance than direct translation methods.

3.2 Machine Learning-Based Approaches

The development of Named Entity Recognition (NER) systems underwent a significant transformation with the introduction of Machine Learning (ML), particularly for morphologically rich and low-resource languages. ML techniques enable greater flexibility and scalability by introducing data-driven methods that learn patterns from annotated corpora, in contrast to rule-based systems, which rely on expert-driven lexicons and handcrafted linguistic rules. This change enabled NER systems to outperform their rule-based predecessors in handling ambiguity, linguistic variation, and domain transfer.

Several early and promising ML techniques were explored across languages in the Sino-Tibetan family. Among the most popular were Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (ME), and Hidden Markov Models (HMM). These models learn patterns from annotated datasets with a specified set of features like word prefixes, suffixes, PoS tags, and context windows. When Feng et al. [14] used the HMM, ME, and CRF models on the Chinese 863 NER task, they found that CRF performed best, outperforming both HMM and ME in terms of accuracy and scalability, with F1-scores of 84.39% and 80.68% on simplified and traditional Chinese, respectively. Similarly, Lay et al. [25] used a small annotated corpus to implement an HMM-based NER system in the Myanmar language. The system showed promising outcomes despite resource constraints, highlighting the feasibility of HMM in low-resource settings. In addition, Hussain et al. [19], built an SVM-based NER system for the Mising language based on a 50,000 word annotated corpus with an F1 score of 87.77%. Similarly, Manipuri systems based on CRF also performed well when feature engineering was carefully managed [44]. However, the adoption of ML approaches has generated a new requirement for annotated datasets. ML-based systems mainly rely on labeled training data, whereas rule-based systems rely on linguistic knowledge. In environments with limited resources, like many Sino-Tibetan languages, where annotated corpora are either nonexistent or very rare, this dependency presents a major bottleneck. Nevertheless, ML-based approaches set the stage for further developments in NER, offering a bridge between rule-based systems and more complex deep learning models. Their adaptability and comparatively lower computational requirements make them a continued area of interest, particularly in situations where deep learning is impractical.

3.3 Deep Learning-Based Approaches

Deep learning has revolutionized NER by automating feature extraction and modeling long-distance dependencies. Architectures like BiLSTM, GRU, CNN, and Transformers dominate modern NER systems. These models learn contextual representations from raw text and often use pre-trained embeddings or language models like BERT, ALBERT, and RoBERTa. For Sino-Tibetan languages, these approaches have demonstrated promising results. A Bodo NER system LSTM model achieved 99.62% accuracy, reflecting its strength in sequence modeling [37]. In Chinese NER, architectures such as the Lattice-Transformer-Graph [28] and CNN-LSTM-CRF [59] push the limits of the F1 scores above 95%. The challenges are abundant despite these advances. It is a major challenge of data scarcity where annotated datasets for languages such as Bodo, Mizo have few. This problem was mitigated by using techniques like data augmentation and transfer learning; in this way, the models become more generalizing to the low-resource language. In 2023, Tang et al. [57] introduced a BERT-Bi-LSTM-AM-CRF model for Chinese NER that combines BERT embeddings with Bi-LSTM layers and attention mechanisms to improve contextual understanding and tagging accuracy; the F1 improvements over MASR and People's Daily datasets are highly significant. Similarly, Hu et al. [18] have proposed a multi-level ALBERT-BiLSTM-CNN-CRF model that is superior to Lattice-LSTM-CRF because it contains the multi-layered contextual embedding mechanism. In clinical NER, An et al. [2] proposed a MUSA-BiLSTM-CRF model with multi-head self-attention to deal with complex Chinese clinical terms to show a significant improvement in benchmark datasets like CCKS. Mo et al. [33] were the first authors who utilized CNN-Bi-LSTM-CRF to adapt a syllable-based Myanmar NER model.

A model architecture proposed achieved an F1 score of 91.3%. Such an architecture captures features unique to Myanmar syllable-based. Jimmy et al. [21] Achieved 98.19% F1 using BiLSTM-CRF with character-level embeddings for Manipuri, which showed the power of character plus word embeddings in rich morphological variation languages. Despite their efficacy, Deep learning models are data-hungry and computationally intensive, making them less feasible for extremely low-resource languages without further adaptation. The Bi-LSTM-CRF model is found to be extremely effective model for NER in Sino-Tibetan languages. It captures the context of past and future words as it has bidirectional processing and makes use of two types of embeddings: character-level and word-level [37]. These will reduce morphological variations and generalize more for rare words.

A CRF layer when added, enforces valid label sequences, so that the model optimizes its overall prediction accuracy. The combination is seen to be the most accurate compared to all the other models in NER for Sino-Tibetan languages. Fig. 1 shows the general architecture of NER using Deep Learning.

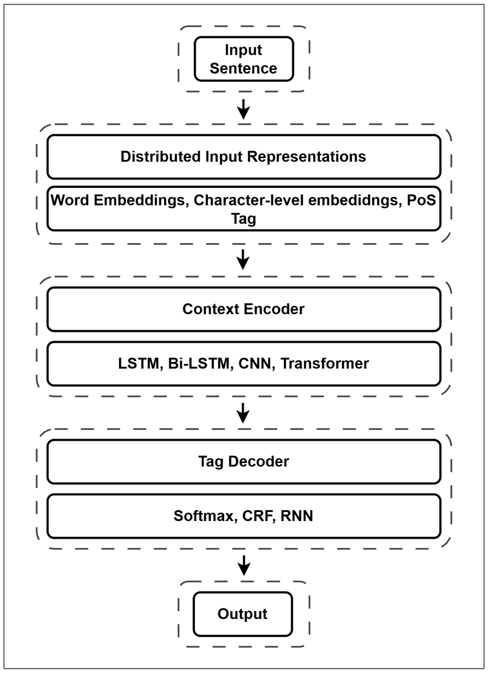
Input Sentence
Distributed Input Representations
Word Embeddings, Character-level embedidngs, PoS Tag
Context Encoder
LSTM, Bi-LSTM, CNN, Transformer
Tag Decoder
Softmax, CRF, RNN
Output

Fig. 1. General architecture of NER using Deep Learning

3.4 Hybrid Approaches

Hybrid systems for Named Entity Recognition (NER) have evolved from initial integrations of rule-based and statistical approaches to more sophisticated architectures that blend conventional linguistic expertise with contemporary machine learning techniques, including deep learning. These approaches combine the strengths of multiple techniques to attain superior performance compared to any singular approach. These methods are particularly appropriate for languages with moderate resources where domain expertise and linguistic intuition can significantly enhance data-driven learning. An early instance is the semi-hybrid NER system for Nepali developed by Dey et al. [12] which integrated rule-based techniques like gazetteer lookups and handcrafted rules with a statistical Hidden Markov Model (HMM). The approach involved the creation of a stemming module, a PoS tagger, and a rule-enhanced HMM-based NER module. This system achieved promising results in identifying entities like persons, locations, and organizations, illustrating how hybrid systems can compensate for limited annotated resources through the incorporation of linguistic structure. Another noteworthy contribution is by Kaur et al. [22] developed a hybrid NER model for Manipuri by combining CRF with heuristic rules, achieving an impressive F1 score of 93.3% even with limited annotated data. A more advanced hybrid system was put forward by Jia et al. [20] for Tibetan person name identification, which combined a Maximum Entropy model with Conditional Random Fields. The system uses boundary word information, name dictionaries, and syntactic information to recognize Tibetan names, which is a task complicated by the absence of word boundaries, heavy use of common nouns as names, and extremely variable length of names.

The system performed better (F1 $\approx$ 93.05%)than either model alone when the outputs of the two models were combined linearly. Better generalization was made possible by this combination, which lessened the shortcomings of both models:

the high precision of CRF and the high recall of the Maximum Entropy model. Such methods show how linguistic patterns like suffix lists or syntactic hints can be embedded into learning structures to enhance entity boundary and type recognition. Nevertheless, even though they are useful, hybrid models tend to entail sophisticated design decisions and need meticulous calibration to maintain uniform performance over language varieties. They need careful integration of rule-based modules with machine learning components so as not to introduce redundancy, and their performance could vary based on the uniformity of language use across domains or dialects.

3.5 Cross-Lingual and Multilingual Approaches

Cross-lingual techniques use data and models from high-resource languages English or Chinese, to assist named entity recognition (NER) in low-resource languages like those in the Sino-Tibetan branch. This method presents a unique strategy because of its orientation towards multilingual knowledge transfer as well as low-resource adaptation. Such methods typically employ multilingual pre-trained models (e.g., mBERT, XLM-R) or bilingual word embeddings to develop shared semantic representations that support tag transfer between languages. For example, Yu et al. [31] employed Tibetan-Chinese cross-lingual embeddings with a BiLSTM-CRF model for enhancing Tibetan location name recognition. Zhu et al. [62] presented the CAN-NER model that made use of a Convolutional Attention Network without lexicons and achieved competitive performance on Chinese corpora. A recent experiment of Bodo NER and PoS tagging evaluated Google's Gemini 2.0 Flash Thinking Experiment model's zero-shot cross-lingual transferability [38]. Both translation-based English-to-Bodo tag projection and prompt-based approach using parallel English-Bodo sentence pairs were tested. The latter proved to be better, particularly for NER, as it facilitated stronger contextual anchoring and more successful tag transfer. Despite their potential, cross-lingual models continue to be confronted by perennial challenges from translation mistakes, grammar divergences, ill-aligned multi-word entities, and semantic mismatches, particularly for morphologically complex languages with scarce resources. However, these techniques offer a promising basis for bootstrapping NER in low-resource environments and indicate the necessity for emerging innovations like prompt engineering, few-shot fine-tuning, and community-driven resource creation.

4 Performance Evaluation Criteria for Named Entity Recognition (NER)

Evaluation metrics are essential to quantify the efficacy and reliability of Named Entity Recognition (NER) systems, offering insights into model performance on different datasets and languages.

For Sino-Tibetan languages, where linguistic features differ drastically, standardized evaluation criteria assist in benchmarking performance and guide improvements.

4.1 Evaluation Metrics

Evaluating Named Entity Recognition (NER) systems generally relies on traditional classification metrics:

— Precision: The ratio of accurately predicted named entities to the total of predicted entities. It describes the accuracy of the model in identifying proper entities.

$$Precision = \frac{TP}{TP + FP}. \tag{1}$$

— Recall: The ratio of accurately predicted named entities to the true amount of entities in the corpus. It shows the success of the model in identifying all proper entities.

$$Recall = \frac{TP}{TP + FN}. \tag{2}$$

— F1-Score: The harmonic mean between precision and recall, a single number balancing between the two.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3}$$

Table 3. Merits and Demerits of NER approaches for Sino-Tibetan languages

Approach	Merits	Demerits
Rule-based approach	- effective for low-resource languages - incorporates linguistic insights specific to each language - simple to interpret and implement	- requires extensive manual rule creation - struggles with language ambiguity and exceptions - poor scalability across domains and dialects
Machine learning approach	- learns patterns from annotated data - generally outperforms rule-based methods - adaptable with adequate training data	- requires significant annotated data, often scarce - time-consuming feature engineering - limited performance on rare entities or OOV words
Deep learning approach	- captures rich contextual information - achieves state-of-the-art performance - leverages pre-trained models for better adaptation	- data and computation-intensive - complex and challenging to interpret - pre-trained models may not be available for all languages
Hybrid approach	- utilizes both rule-based and data-driven learning - effective in low-resource settings - adaptable to specific language needs	- complex design and implementation - requires careful balance of rule and data-driven components - performance varies by language and dependent on effective component integration
Cross-lingual approach	- leverages high-resource languages like Chinese to aid low-resource languages - effective for sharing representations across languages - beneficial for low-resource language adaptation	- language-specific nuances may not transfer well - performance relies on similarity between source and target languages - requires careful source selection

Here:

— TP (True Positives): Correctly predicted named entities.

— FP (False Positives): Incorrectly predicted entities (not present or misclassified).

— FN (False Negatives): Entities that are in the ground truth but not detected by the model.

These metrics were standardized in cooperative tasks such as CoNLL-2003 [50], which predominantly are used by NER research on Sino-Tibetan languages such as research on Tibetan, Burmese, and Chinese dialects.

Such standardized metrics may, however, not capture the semantic accuracy of prediction or entirely capture partially matched entity boundaries, a factor most important in morphologically dense languages.

4.1.1 Aggregation Methods

- Macro-average: This method computes metrics independently for each class and averages the results with equal weight to each class irrespective of how many instances it has. This approach is handy when measuring performance against different types of entities, particularly in covering how well the model can deal with minority classes.

- Micro-average: This is a method that aggregates predictions over all classes before calculating metrics, with equal weight to each single instance irrespective of its class. This approach is appropriate for imbalanced datasets where overall performance on all instances takes precedence.

4.2 Entity-Level and Token-Level Evaluation

Evaluation can be conducted at:

- Entity Level: Demands a strict match of both entity boundaries as well as types. It is more restrictive and is employed in the standards of the majority of benchmarks (e.g., CoNLL).

- Token Level: Evaluates correctness at the token level, beneficial for the study of partial recognition and boundary mismatches, primarily suitable for morphologically rich languages with complex tokenization.

4.3 Exact Match and Partial Match Metrics

NER evaluation in morphologically intricate Sino-Tibetan languages tends to be aided by partial match metrics (e.g., MUC-6 scoring [21]), which award partial credit for predictions with overlapping or partially accurate entity boundaries. This is useful in languages such as Tibetan and Burmese, where boundary uncertainty is typical owing to agglutinative morphology.

Standard benchmarking datasets such as those given by MUC-6 [11], ACE [13], and CoNLL shared tasks [50] have shaped evaluation practices significantly. Recent surveys and reviews (e.g., IEEE TKDE 2020 [26]) shows that these standard practices to Sino-Tibetan languages need to take careful note of language-specific characteristics like tokenization, boundary uncertainty, and annotation conventions.

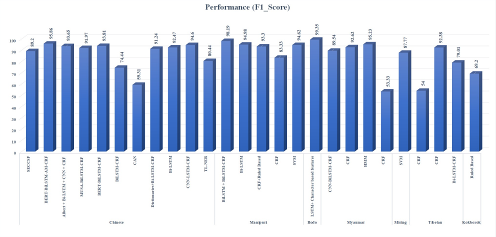

Fig. 2. F-Score Analysis of NER Models for Sino-Tibetan Languages

5 Analysis

In this section, we conduct a close comparative examination of the existing techniques used in Named Entity Recognition (NER) for Sino-Tibetan languages with regard to their adaptability and performance in low-resource settings.

Table 4 illustrates a number of approaches with datasets, target languages, and performance metrics in the form of F-scores, recall, and precision placed in a systematic manner to have a proper perception of the strengths and weak points of each approach.

This table presents how models, for instance, Bi-LSTM-CRF, CNN, and hybrid architectures are being utilized to boost the performance of NER across all limited-resource languages. Figure 2 provides a graphical representation for the F-scores of different NER models for Sino-Tibetan languages to enable making fast visual comparisons on how these models handle complexity due to languages.

This analysis both points out progress in model accuracy and shows the challenges that remain, such as increasing annotated datasets and adaptable architectures suitable for morphologically rich languages from the Sino-Tibetan family.

Table 4. Comparison of NER Methods, Datasets, and Performances for Sino-Tibetan Languages

Reference	Authors & Year	Language	Dataset	Methods	Performance
[27]	Liu et al. (2024)	Chinese	EPIC: Epidemiological Investigation of COVID-19 dataset for CNER	Three-stage annotation method and SECCSF method for CNER	F1 score: 89.2%
[37]	Narzary et al. (2024)	Bodo	Bodo PoS dataset (TDIL)	LSTM + Character-level features	Acc: 99.62%; Prec: 99.75%; Rec: 98.74%; F1: 99.35%
[57]	Tang et al. (2023)	Chinese	MASR and People's Daily datasets	Multi-task BERT-BiLSTM-AM-CRF	F1: 95.86%; Recall: 96.01%; Precision: 95.73%
[21]	Jimmy et al. (2023)	Manipuri	Manipuri NER Corpus	BiLSTM + CRF + Word / Char Embeddings	F1: 98.19%; Cluster Acc: 88.14%
[18]	S et al. (2023)	Chinese	Real datasets	Albert + Bi-LSTM + CNN + CRF	F1: 93.65%
[2]	An et al. (2022)	Chinese	CCKS2017-CNER & CCKS2018-CNER	MUSA-BiLSTM-CRF	F1: 91.97 (2017); 91.81 (2018)
[17]	B et al. (2021)	Chinese	Network Security Domain	BERT-BiLSTM-CRF	F1: 93.81%
[61]	Q et al. (2020)	Chinese	OntoNotes4, Weibo NER, Chinese Resume	BiLSTM-CRF + Attention	F1: 74.44% (ON); 59.63% (Weibo); 95.58% (Resume)
[33]	Mo et al. (2020)	Myanmar	170k NEs, 60k sentences	CNN-BiLSTM-CRF	F1: 89.54% (10-fold CV)
[62]	Zhu et al. (2019)	Chinese	Weibo, Resume, OntoNotes, MSRA	CAN model	F1: 59.31–94.94%
[58]	Wang et al. (2019)	Chinese	CCKS-2017 Task 2	Dictionaries + BiLSTM-CRF	F1: 91.24%; Prec: 90.83%; Rec: 91.64%
[47]	Ouyang et al. (2017)	Chinese	People's Daily	Bi-LSTM + Word Embedding	F1: 92.47%
[24]	Laishram et al. (2020)	Manipuri	Sanghai Express	Bi-LSTM + Word Embedding + PoS Clusters	F1 (MNE): 94.98%
[19]	Hussain et al. (2016)	Mising	50K word corpus	SVM (12 tags)	F1: 87.77%; Rec: 90.58%; Prec: 85.14%
[22]	Kaur et al. (2013)	Manipuri	Manipuri Document	CRF + Gazetteer Rules	F1: 93.3%; Rec: 92.26%; Prec: 94.27%
[44]	Nongmeikapam et al. (2011)	Manipuri	Manipuri Corpus	CRF	F1: 83.33%; Rec: 81.12%; Prec: 85.67%
[48]	Patra et al. (2015)	Kokborok	30k words (Bible, news)	Rule-based+Supervised (MIRA)	F1: 69.2% (Rule); 83.18% (Sup)
[29]	Liu et al. (2018)	Tibetan	249 training sentences	CRF + Active Learning	F1: 10.7→54 (Conf); 46.5 (NE-feat)
[36]	Nandar et al. (2020)	Myanmar	ALT Parallel Corpus	CRF	F1: 92.62% (syll); 84.88% (char)
[25]	Lay et al. (2019)	Myanmar	Raw Corpus	HMM	F1: 97.21%; Acc: 95.23%
[32]	Mo et al. (2017)	Myanmar	News Articles	CRF	Prec: 50%; Rec: 57.14%; F1: 53.33%
[59]	Wu et al. (2019)	Chinese	MSRA	CNN-LSTM-CRF	F1: 94.6% (B3); 93.9% (B4)
[49]	Peng et al. (2020)	Chinese	Legal, OntoNotes, MSRA	TL-NER	F1: 80.44% (Legal); 91.42% (MSRA)
[20]	Jia et al. (2016)	Tibetan	Tibet Daily	CRF & MaxEnt	F1: 92.38% (CRF); 91.55% (ME)
[54]	Singh et al. (2010)	Manipuri	Web Corpus	SVM	F1: 94.07% (RMWE); 93.96% (MNE)
[31]	Ma et al. (2019)	Tibetan	Location Name Dataset	Bi-LSTM-CRF	F1: 79.01%

6 Issues and Challenges of NER in Sino-Tibetan languages

The following issues and challenges of NER in Sino-Tibetan languages are identified.

- Lack of resources: Sino-Tibetan languages are often low-resource and under-resourced languages, which means there is a scarcity of annotated data, linguistic tools, and research for NER. Annotation of data manually is costly and time-consuming and may require the collaboration of native speakers and linguistic experts [54].
- Linguistic complexity: Sino-Tibetan languages have already contributed a great amount of linguistic complexity to orthography, morphology, syntax, and semantics. It is tough to identify the word segmentation, named entity boundary detection, and classification of named entity types in these languages. In some languages, the morphology is rich and agglutinative, which may generate lots of variants of the same entity. In some languages, the entity classes may be ambiguous and inconsistent; therefore, more fine-grained and normalized annotation might be required.
- Cultural diversity: Sino-Tibetan languages process varying scripts, dialects, and cultural features. These affect the recognition and categorization of named entities. For instance, there are few languages that lack word boundaries, capitalization, and punctuation characters; hence, it might be difficult to detect the entity boundaries. There are languages with varying naming conventions, honorifics, or transliterations that can impact the entity recognition and normalization of named entities. Certain languages possess various types and domains of named entities, which can involve additional domain-specific knowledge and resources.
- Named entity schemes and categories: Sino-Tibetan languages possess a diversified and extensive range of named entities, which may not be represented by existing NER schemes or categories.
- Domain Adaptation: Sino-Tibetan texts span across various domains, such as news, literature, religion, and social media, networks having different styles, vocabularies, and named entity types [3]. There could be different domain-specific rules for different languages for NER, which may be difficult due to their unique linguistic features. The complexity of the language makes it challenging to create effective handcrafted rules for entity recognition [21].
- Multilingualism: Some texts like Manipuri texts often contain words or names from other languages, such as Bengali, which require different recognition methods and resources [48].
- Evaluation of methods and resources: No well-defined evaluation metrics, datasets, baselines, and benchmarks are available regarding NER in Sino-Tibetan languages, due to which a comparison between the effectiveness and efficiency evaluation of different methods and resources becomes hard to realize [48].
- Ethical and social issues: NER for Sino-Tibetan languages may face some ethical and social problems, such as privacy and security of the named entities, biasness of the methods and resources, impact or responsibility of the NLP applications, etc.

7 Conclusion and Future Work

The reviewed works highlight significant progress in applying NER techniques to Sino-Tibetan languages, especially through deep learning models such as Bi-LSTM-CRF, attention mechanisms, and hybrid approaches. Such methods are effective in capturing contextual cues and label dependencies, though challenges remain, such as limited annotated data and high computational demands. That is, deep learning models have high accuracy, but resource requirements limit their applicability. In contrast, machine learning methods like CRF and SVM offer a practical balance between performance and resource requirements, and rule-based systems remain valuable when linguistic rules and gazetteers are available. The future of NER research prioritize the development of resource-efficient, adaptable models, explores robust data augmentation strategies, and leverages transfer learning and multilingual training for better cross-lingual generalization.

Our focal point in our research continuum lies in creating large, annotated datasets for Bodo, a lesser-resourced Sino-Tibetan language. An equally promising future research direction might be the possibility of transfer learning from related richer-resource languages to further enhance NER capabilities across the Sino-Tibetan language family.

References

1. **Ahmed, M., Talukdar, K., Boruah, P., Sarma, S. K., Kashyap, K. (2023).** Guit-nlp's submission to shared task: Low resource indic language translation. Proceedings of the Eighth Conference on Machine Translation, pp. 935–940.

2. **An, Y., Xia, X., Chen, X., Wu, F.-X., Wang, J. (2022).** Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. Artificial Intelligence in Medicine, Vol. 127, pp. 102282.

3. **Baruah, K. K., Das, P., Hannan, A., Sarma, S. K. (2014).** Assamese-english bilingual machine translation. arXiv preprint arXiv:1407.2019.

4. **Baruah, N., Gogoi, A., Neog, M. (2023).** Detection of hate speech in assamese text. International Conference on Communication and Computational Technologies, Springer Nature Singapore, pp. 655–670. DOI: 10.1007/978-981-99-3485-0$_52$.

5. **Baruah, N., Sarma, S. K., Borkotokey, S. (2019).** A novel approach of text summarization using assamese wordnet. 2019 4th international conference on information systems and computer networks (ISCON), IEEE, pp. 305–310.

6. **Baruah, N., Sarma, S. K., Borkotokey, S. (2021).** A single document assamese text summarization using a combination of statistical features and assamese wordnet. Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2019, Volume 2, Springer Singapore, pp. 125–136. DOI: 10.1007/978-981-15-4032-5$_12$.

7. **Baruah, N., Sarma, S. K., Borkotokey, S., Borah, R., Phukan, R. D., Gogoi, A. (2022).** A graph-based extractive assamese text summarization. Computational Methods and Data Engineering: Proceedings of ICCMDE 2021, Springer Nature Singapore, pp. 1–12. DOI: 10.1007/978-981-16-9275-4$_1$.

8. **Basumatary, B., Rahman, M., Sarma, S. K., Boruah, P. A., Talukdar, K. (2023).** Deep learning based bodo parts of speech tagger. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, pp. 1–5.

9. **Bentham, J., Pakray, P., Majumder, G., Lalbiaknia, S., Gelbukh, A. (2016).** Identification of rules for recognition of named entity classes in mizo language. 2016 Fifteenth Mexican International Conference on Artificial Intelligence (MICAI), IEEE, pp. 8–13.

10. **Brahma, B., Barman, A., Sarma, S. K., Boro, B. (2012).** Corpus building of literary lesser rich language-bodo: Insights and challenges. Proceedings of the 10th Workshop on Asian Language Resources, pp. 29–34.

11. **Chinchor, N., Robinson, P. (1997).** Muc-7 named entity task definition. Proceedings of the 7th Conference on Message Understanding, Vol. 29, pp. 1–21.

12. **Dey, A., Paul, A., Purkayastha, B. S. (2014).** Named entity recognition for nepali language: A semi hybrid approach. International Journal of Engineering and Innovative Technology (IJEIT) Volume, Vol. 3, pp. 21–25.

13. **Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M.,**

Weischedel, R. M., et al. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. Lrec, Citeseer, Vol. 2, pp. 837–840.

14. **Feng, Y., Sun, L., Zhang, J. (2005).** Early results for chinese named entity recognition using conditional random fields model, hmm and maximum entropy. 2005 International Conference on Natural Language Processing and Knowledge Engineering, IEEE, pp. 549–552.

15. **Gogoi, A., Baruah, N. (2022).** A lemmatizer for low-resource languages: Wsd and its role in the assamese language. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 21, No. 4, pp. 1–22. DOI: 10.1145/3502157.

16. **Goutom, P. J., Baruah, N., Sonowal, P. (2023).** An abstractive text summarization using deep learning in assamese. International Journal of Information Technology, Vol. 15, No. 5, pp. 2365–2372. DOI: 10.1007/s41870-023-01279-7.

17. **He, B., Chen, J. (2021).** Named entity recognition method in network security domain based on bert-bilstm-crf. 2021 IEEE 21st International Conference on Communication Technology (ICCT), IEEE, pp. 508–512.

18. **Hu, S., Guan, J., Li, W. (2022).** Chinese named entity recognition based on multi-level information extraction. 2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI), IEEE, pp. 554–557.

19. **Hussain, S., Kuli, J. J., Hazarika, G. C. (2016).** The first step towards named entity recognition in mising language. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, pp. 3013–3016.

20. **Jia, Y., Li, Y., Zong, C., Yu, H. (2016).** A hybrid approach using maximum entropy model and conditional random fields to identify tibetan person names. Himalayan Linguistics, Vol. 15, No. 1.

21. **Jimmy, L., Nongmeikappam, K., Naskar, S. K. (2023).** Bilstm-crf manipuri ner with character-level word representation. Arabian Journal for Science and Engineering, Vol. 48, No. 2, pp. 1715–1734.

22. **Kaur, D. (2013).** Named entity recognition in manipuri: a hybrid approach. Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25-27, 2013. Proceedings, Springer, pp. 104–110.

23. **Khargharia, D., Newar, N., Baruah, N. (2018).** Applications of text summarization.. International Journal of Advanced Research in Computer Science, Vol. 9, No. 3.

24. **Laishram, J., Nongmeikapam, K., Naskar, S. K. (2020).** Deep neural model for manipuri multiword named entity recognition with unsupervised cluster feature. Proceedings of the 17th International Conference on Natural Language Processing (ICON), pp. 420–429.

25. **Lay, K. K., Cho, A. (2019).** Myanmar named entity recognition with hidden markov model. International Journal of Trend in Scientific Research and Development (IJTSRD), Vol. 3, No. 4, pp. 1144–1147.

26. **Li, J., Sun, A., Han, J., Li, C. (2020).** A survey on deep learning for named entity recognition. IEEE transactions on knowledge and data engineering, Vol. 34, No. 1, pp. 50–70.

27. **Li, P., Zhou, G., Guo, Y., Zhang, S., Jiang, Y., Tang, Y. (2024).** Epic: An epidemiological investigation of covid-19 dataset for chinese named entity recognition. Information Processing & Management, Vol. 61, No. 1, pp. 103541.

28. **Lin, M., Xu, Y., Cai, C., Ke, D., Su, K. (2023).** A lattice-transformer-graph deep learning model for chinese named entity recognition. Journal of Intelligent Systems, Vol. 32, No. 1, pp. 20222014.

29. **Liu, F. F., Wang, Z. J. (2018).** Active learning for tibetan named entity recognition based on crf. LREC 2018 Workshop, pp. 18.

30. **Liu, S., Best, M. (2025).** A survey of nlp progress in sino-tibetan low-resource languages. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 7804–7825.

31. **Ma, W., Zhao, K. (2019).** Tibetan location name recognition using tibetan-chinese cross-lingual word embeddings. 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), IEEE, pp. 384–387. DOI: 10.1109/MLBDBI48998.2019.00082.

32. **Mo, H. M., Nwet, K. T., Soe, K. M. (2017).** Crf-based named entity recognition for myanmar language. Genetic and Evolutionary Computing: Proceedings of the Tenth International Conference on Genetic and Evolutionary Computing, Springer, pp. 204–211.

33. **Mo, H. M., Soe, K. M. (2020).** Myanmar named entity corpus and its use in syllable-based neural named entity recognition. International Journal of Electrical and Computer Engineering, Vol. 10, No. 2, pp. 1544–1551.

34. **Moirangthem, G., Nongbri, L., Johny Singh, N., Nongmeikapam, K. (2022).** Embeddings-based parallel corpus creation for english-manipuri. International Conference on Communication and Intelligent Systems, Springer, pp. 489–502.

35. **Moirangthem, G., Nongmeikapam, K. (2021).** A back-transliteration based manipuri meetei mayek keyboard ime. 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, pp. 1–6.

36. **Nandar, T. L., Soe, T. L., Soe, K. M. (2020).** A comparative study of named entity recognition on myanmar language. 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, pp. 60–64.

37. **Narzary, S., Brahma, A., Nandi, S., Som, B. (2024).** Deep learning based named entity recognition for the bodo language. Procedia Computer Science, Vol. 235, pp. 2405–2421.

38. **Narzary, S., Brahma, B., Mahilary, H., Brahma, M., Som, B., Nandi, S. (2025).** Comparative study of zero-shot cross-lingual transfer for bodo pos and ner tagging using gemini 2.0 flash thinking experimental model. arXiv preprint arXiv:2503.04405.

39. **Neog, M., Baruah, N. (2023).** A deep learning framework for assamese toxic comment detection: Leveraging lstm and bilstm models with attention mechanism. International Conference on Advances in Data-driven Computing and Intelligent Systems, Springer Nature Singapore, pp. 485–497. DOI: 10.1007/978-981-99-3485-0_45.

40. **Nongmeikapam, K., Bandyopadhyay, S. (2011).** Genetic algorithm (ga) in feature selection for crf based manipuri multiword expression (mwe) identification. arXiv preprint arXiv:1111.2399.

41. **Nongmeikapam, K., Laishram, D., Singh, N. B., Chanu, N. M., Bandyopadhyay, S. (2011).** Identification of reduplicated multiword expressions using crf. Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I 12, Springer, pp. 41–51.

42. **Nongmeikapam, K., Nonglenjaoba, L., Nirmal, Y., Bandyopadhyay, S. (2012).** Reduplicated mwe (rmwe) helps in improving the crf based manipuri pos tagger. arXiv preprint arXiv:1203.4933.

43. **Nongmeikapam, K., Nonglenjaoba, L., Roshan, A., Singh, T. S., Singh, T. N., Bandyopadhyay, S. (2012).** Transliterated svm based manipuri pos tagging. Advances in Computer Science, Engineering & Applications: Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), May 25-27, 2012, New Delhi, India, Volume 1, Springer, pp. 989–999.

44. **Nongmeikapam, K., Shangkhunem, T., Chanu, N. M., Singh, L. N., Salam, B., Bandyopadhyay, S. (2011).** Crf based name entity recognition (ner) in manipuri: A highly agglutinative indian language. 2011 2nd National Conference on Emerging Trends and Applications in Computer Science, IEEE, pp. 1–6.

45. **Nongmeikapam, K., Sharma, A. U., Devi, L. M., Keisam, N., Singh, K. D., Bandyaopadhyay, S. (2012).** Will the identification of reduplicated multiword expression (rmwe) improve the performance of svm based manipuri pos tagging? Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I 13, Springer, pp. 117–129.

46. **Nongmeikapam, K., Singh, T. I., Chanu, N. M., Bandyopadhyay, S. (2014).** Manipuri chunking: An incremental model with pos and rmwe. Proceedings of the 11th International Conference on Natural Language Processing, pp. 277–286.

47. **Ouyang, L., Tian, Y., Tang, H., Zhang, B. (2017).** Chinese named entity recognition based on b-lstm neural network with additional features. Security, Privacy, and Anonymity in Computation, Communication, and Storage, Springer, pp. 269–279.

48. **Patra, B. G., Debbarma, N., Das, D., Bandyopadhyay, S. (2015).** Named entity recognizer for less resourced language kokborok. 2015 International Conference on Asian Language Processing (IALP), IEEE, pp. 164–168.

49. **Peng, D., Wang, Y., Liu, C., Chen, Z. (2020).** Tl-ner: A transfer learning model for chinese named entity recognition. Information Systems Frontiers, Vol. 22, No. 6, pp. 1291–1304.

50. **Sang, E. F., De Meulder, F. (2003).** Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

51. **Sarma, S. K., Brahma, B., Gogoi, M., Ramchiary, M. B. (2010).** A wordnet for bodo language: Structure and development. Global Wordnet Conference (GWC10), Mumbai, India, pp. n/a.

52. **Sarma, S. K., Sarmah, D., Brahma, B., Bharali, H., Mahanta, M., Saikia, U. (2012).** Building multilingual lexical resources using wordnets: Structure, design and implementation. Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, pp. 161–170.

53. **Sarmah, J., Sarma, S. K., Barman, A. K. (2019).** Development of assamese rule based stemmer using wordnet. Proceedings of the 10th Global WordNet Conference, pp. 135–139.

54. **Singh, T. D., Bandyopadhyay, S. (2010).** Web based manipuri corpus for multiword ner and reduplicated mwes identification using svm. Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing, pp. 35–42.

55. **Talukdar, K., Sarma, S. K. (2023).** Upos tagger for low resource assamese language: Lstm and bilstm based modelling. 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), IEEE, pp. 1–6.

56. **Talukdar, K., Sarma, S. K., Naznin, F., Kashyap, K. (2023).** Influence of data quality and quantity on assamese-bodo neural machine translation. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, pp. 1–5.

57. **Tang, X., Huang, Y., Xia, M., Long, C. (2023).** A multi-task bert-bilstm-am-crf strategy for chinese named entity recognition. Neural Processing Letters, Vol. 55, No. 2, pp. 1209–1229.

58. **Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., He, P. (2019).** Incorporating dictionaries into deep neural networks for the chinese

clinical named entity recognition. Journal of Biomedical Informatics, Vol. 92, pp. 103133.

59. **Wu, F., Liu, J., Wu, C., Huang, Y., Xie, X. (2019).** Neural chinese named entity recognition via cnn-lstm-crf and joint training with word segmentation. The World Wide Web Conference, pp. 3342–3348.

60. **Yu, H., Jiang, T., Ma, N. (2010).** Named entity recognition for tibetan texts using case-auxiliary grammars. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1, pp. n/a.

61. **Zhong, Q., Tang, Y. (2020).** An attention-based bilstm-crf for chinese named entity recognition. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), IEEE, pp. 550–555.

62. **Zhu, Y., Wang, G., Karlsson, B. F. (2019).** Can-ner: Convolutional attention network for chinese named entity recognition. arXiv preprint arXiv:1904.02141.

*Article received on 30/05/2025; accepted on 27/07/2025. *Corresponding author is Jinia Angeline Gayary.*