

NerAxom-MGF: Leveraging Multi-Granularity Fusion for Assamese Named Entity Recognition

Punam Sarmah

*Department of Computer Science and Engineering
Tezpur University Napaam, Tezpur India*

PUSARMAH@GMAIL.COM

Manash P Lahkar

*Department of Computer Science and Engineering
Tezpur University Napaam, Tezpur India*

CSP21104@TEZU.AC.IN

Shobhanjana Kalita

*Department of Computer Science and Engineering
Tezpur University Napaam, Tezpur India*

KALITAS@TEZU.ERNET.IN

Utpal Sharma

*Department of Computer Science and Engineering
Tezpur University Napaam, Tezpur India*

UTPAL@TEZU.ERNET.IN

Editor: Debarshi Kumar Sanyal

Abstract

We introduce **MGF-NER**, a multi-granularity fusion architecture that integrates character-level, subword-level, and entity-span-level representations for the same task using **NerAxom** dataset. NerAxom comprises 102,857 tokens across 4,178 sentences annotated under the BIO scheme with seven entity categories: Person, Location, Organization, Date, Number, Occupation, and Work-of-Art. MGF-NER combines a Character-level Convolutional Neural Network that captures morpheme-level cues attenuated by subword fragmentation, a pre-trained MuRIL transformer for cross-lingually pretrained contextual semantics, an Entity Span Attention module for multi-word boundary detection, and a Conditional Random Field decoder that enforces valid BIO transitions. Under a seeded five-run evaluation protocol on the NerAxom test set, MGF-NER achieves a mean micro-averaged F1 of 85.45 ± 0.46 , outperforming the strongest single-encoder baseline MuRIL (83.88 ± 0.79) by 1.57 F1 points; the improvement is statistically significant under a paired sentence-level Wilcoxon signed-rank test ($p = 0.005$). We benchmark MGF-NER against nine baselines spanning traditional feature-based sequence labelling (CRF), a recurrent baseline (BiLSTM-CRF), language-agnostic multilingual transformers (mBERT, XLM-RoBERTa base and large, DistilBERT-multi), a script-similar Bangla-specialised transformer (Bangla-BERT), and Indic-specialised transformers (IndicBERTv2). Single-component and compound-component ablations, together with per-category ablation, decompose the gain across the four architectural components, and error analysis characterises the residual error profile.

Keywords: Named Entity Recognition, Assamese, Low-Resource NLP, Multi-Granularity Fusion

1 Introduction

Named Entity Recognition (NER), the task of identifying and classifying named entities such as persons, locations, and organisations in text, is a foundational component of information extraction, question answering, machine translation, and knowledge graph construction (Nadeau and Sekine, 2007). In information retrieval specifically, NER supports semantic search by moving retrieval systems beyond surface-level keyword matching: entity-aware document and query representations improve ranking, entity-based search, and question answering by resolving surface variants of the same entity and grouping them under a single semantic anchor (Nadeau

and Sekine, 2007). For low-resource languages, this role is even more consequential. Retrieval systems for such languages routinely suffer from limited vocabulary coverage and inconsistent surface representations of the same entity, and NER mitigates both by highlighting semantically important units within the text and reducing dependence on large surface-form vocabularies (Peters et al., 2018).

Progress on NER for resource-rich languages has been substantial, with state-of-the-art English systems reaching F1 scores of 93–94 on CoNLL-2003 (Devlin et al., 2019; Straková et al., 2019). For low-resource languages, the picture is markedly different: annotated data is scarce, pretrained language models cover the language only marginally, and the linguistic characteristics of the target language may not be well represented in general-purpose multilingual encoders (Hedderich et al., 2021). Assamese (অসমীয়া), an Eastern Indo-Aryan language spoken by approximately 15 million people predominantly in the Indian state of Assam, is a canonical low-resource language in this sense. It has (i) a productive inflectional and derivational morphology, (ii) a Brahmi-descended script that shares most of its inventory with Bengali but includes distinct characters for the /r/ and /w/-/v/ sounds, (iii) relatively free word order that permits multiple valid arrangements of the same clause, and (iv) a high prevalence of multi-word named entities, particularly for organisations, creative works, and honorific-prefixed personal names (Sharma et al., 2010).

These characteristics complicate NER in three concrete ways that motivate our design. First, productive suffixation attaches case markers, honorifics, and quantifiers directly to entity heads, so morpheme-level information becomes category-diagnostic. For example, the locative case marker -ত attaches to the place name কাজিৰঙা (*Kaziranga*) to form কাজিৰঙাত (*at Kaziranga*), which is a single token in the surface string but a two-morpheme unit whose head is a location entity. Similarly, the honorific ডঃ (*Dr.*) preceding a name forms an entity mention that spans across a discourse-level marker; conflating or splitting it changes the entity boundary. Second, multi-word institutional names of the form তেজপুৰ বিশ্ববিদ্যালয় (*Tezpur University*) or অসম গ্ৰামীণ বিকাশ বেংক (*Assam Gramin Vikash Bank*) require the model to recognise multi-token span structure, not merely to classify individual tokens. Third, and empirically, when general-purpose multilingual subword tokenisers are applied to Assamese, they frequently fragment inflected and compound forms in ways that split across morpheme boundaries. As we illustrate in Section 4.3 (Table 6), the compound বিশ্ববিদ্যালয় (*university*) is split by MuRIL’s WordPiece tokeniser into pieces that do not correspond to Assamese morphemes, and the locative form কাজিৰঙাত is split into three pieces that separate the case marker from its head. This fragmentation attenuates morpheme-level signals for category discrimination, which motivates supplementing the subword backbone with a character-level encoder that operates strictly below the WordPiece level.

Existing NER systems for Assamese have predominantly relied on rule-based approaches or on statistical models with handcrafted features (Sharma et al., 2010, 2014; Talukdar et al., 2014). The performance of these systems, typically in the 60–65 F1 range for the principal categories, is insufficient for practical downstream applications and is not competitive with what pretrained multilingual encoders now achieve on other Indic languages. The emergence of Indic-specialised pretrained models, in particular MuRIL (Khanuja et al., 2021), has opened new avenues, but as our results in Section 6 show, directly applying MuRIL to Assamese NER without language-specific adaptations leaves substantial performance on the table: MuRIL alone reaches a mean F1 of 83.88 on the NerAxom corpus, whereas the same backbone augmented with Assamese-specific components reaches 85.45, a difference that is significant under a paired sentence-level test ($p = 0.005$).

We address these gaps in this work with Multi-Granularity Fusion NER (MGF-NER), a neural architecture that integrates three levels of representational granularity for Assamese NER: a Character-CNN that captures morpheme-level cues attenuated by subword fragmentation, MuRIL for cross-lingually pretrained contextual semantics, and an Entity Span Attention module that explicitly models multi-word entity boundaries. A CRF decoder on top of these

representations enforces valid BIO tag transitions, so predicted sequences are structurally consistent by construction. The design principle behind MGF-NER is straightforward: no single representational granularity is sufficient for morphologically rich, low-resource NER, but a small number of carefully chosen granularities, combined with a structured decoder, are sufficient to close a meaningful fraction of the gap between the strong single-encoder baseline and the current ceiling on this corpus.

The main contributions of this paper are as follows:

1. **Architecture.** We propose MGF-NER, a multi-granularity fusion architecture that combines Char-CNN morphological features, MuRIL contextual representations, an Entity Span Attention module, and CRF decoding.
2. **State-of-the-art performance.** Under a seeded five-run evaluation protocol, MGF-NER reaches a mean F1 of 85.45 ± 0.46 on NerAxom, outperforming the strongest single-encoder baseline MuRIL by 1.57 F1 points (paired sentence-level Wilcoxon, $p = 0.005$).
3. **Comprehensive evaluation.** We compare against nine baselines spanning traditional feature-based sequence labelling, deep recurrent architectures, language-agnostic multilingual transformers (mBERT, XLM-RoBERTa base and large, DistilBERT-multi), a script-similar Bangla-specialised transformer (Bangla-BERT), and Indic-specialised transformers (IndicBERTv2, MuRIL). All ten systems are trained under the same seeded protocol and evaluated on the same partition, yielding directly comparable variance estimates and providing a reproducible benchmark for future work.
4. **Detailed analysis.** We report per-category F1, single-component and compound-component ablations, per-category ablation results that localise the contribution of each module, dropout sensitivity, error analysis with a token-level BIO confusion matrix, and a computational cost comparison.

The remainder of this paper is organised as follows. Section 2 reviews related work on NER, on NER for Indian languages including the trajectory of Assamese-specific work, on multi-granularity NER architectures, and on transfer learning for low-resource NER. Section 3 describes the NerAxom corpus, its annotation scheme, and its statistical properties. Section 4 formalises MGF-NER and states the role of each component. Section 5 details the baselines, hyperparameters, seeded protocol, and evaluation metrics. Section 6 presents the main results, per-category performance, ablations, error analysis, and computational efficiency. Section 7 discusses why the multi-granularity design works, the limitations of the current corpus and architecture, and directions for future work.

2 Related Work

This section reviews prior research relevant to the proposed MGF-NER framework. We first discuss the evolution of NER from rule-based approaches to modern neural and transformer-based architectures, followed by studies on NER for Indian and Assamese languages. We then examine multi-granularity representation learning and transfer learning strategies that have been widely adopted for low-resource NER, highlighting the research gaps addressed by the proposed approach.

2.1 Named Entity Recognition: From Rules to Neural Sequence Labelling

NER has progressed through three broad paradigms over the past three decades: rule- and gazetteer-based systems, statistical sequence models with handcrafted features, and neural architectures with learned representations. Early systems relied on handcrafted rules, gazetteers,

and finite-state cascades (Grishman and Sundheim, 1996); while these systems achieved reasonable precision within narrow domains, their recall degraded sharply on out-of-domain text and adapting them to new entity types required substantial manual effort.

The introduction of statistical sequence models replaced hand-authored rules with data-driven inference. Hidden Markov Models (Bikel et al., 1997) and, most influentially, linear-chain Conditional Random Fields (McCallum and Li, 2003) became the standard formulation. CRF-based systems using engineered features (word shape, prefixes and suffixes, part-of-speech tags, gazetteer indicators, and ± 2 -token context) dominated NER benchmarks for over a decade. On the widely used CoNLL-2003 English benchmark, feature-engineered CRF systems reached F1 scores of 88–90 (Ratinov and Roth, 2009), and remain competitive baselines in low-resource settings where learned representations are hard to acquire.

Neural sequence labelling then removed the need for manual feature engineering. Huang et al. (2015) introduced the BiLSTM-CRF architecture, in which a bidirectional LSTM computes per-token contextual features and a linear-chain CRF decodes a globally consistent BIO sequence. On CoNLL-2003 English, this architecture reached 90.10 F1 without gazetteers. Lample et al. (2016) extended it with character-level BiLSTM representations concatenated to word embeddings, reaching 90.94 F1 and establishing the character-word hybrid design that is now standard for morphologically rich languages. Ma and Hovy (2016) showed that the character-level component can equivalently be implemented as a convolutional network (Char-CNN) with essentially matched performance, reaching 91.21 F1 while being simpler to train, and this Char-CNN variant is the direct antecedent of the character encoder in MGF-NER. Dos Santos and Guimaraes (2015) had earlier shown that adding character-level CNN features to a window-based tagger for Portuguese and Spanish yielded consistent gains, particularly for out-of-vocabulary named entities, which is the same phenomenon that motivates the character encoder in Assamese.

Pretrained contextual encoders subsequently redefined the state of the art. ELMo (Peters et al., 2018) introduced deep bidirectional LSTM-based contextualised embeddings, pushing CoNLL-2003 English to 92.22 F1. BERT (Devlin et al., 2019) then reached 92.8 F1 with a transformer-based masked language modelling objective, and its successors reached higher still. More recent work has explored span-based formulations (Zhong and Chen, 2021), template-based prompting (Cui et al., 2021), and unified frameworks that jointly handle flat, nested, and discontinuous NER (Wang et al., 2022; Li et al., 2020). Across these developments, the character-word hybrid design of Lample et al. (2016) and Ma and Hovy (2016) has proven durable: contemporary systems for morphologically rich languages routinely augment transformer backbones with character-level signals, and this is the design tradition to which MGF-NER contributes.

2.2 NER for Indian Languages

NER for Indian languages has developed unevenly across resource levels. Hindi has received the most attention, supported by the FIRE shared task series (Abinaya et al., 2014) and by the Naamapadam corpus (Mhaske et al., 2023), which covers 11 Indian languages with silver-standard annotations. Hindi NER systems combining BiLSTM-CRF with FastText and multilingual BERT reach F1 scores in the range of 78–82 on FIRE-derived splits (Bhattacharjee et al., 2019; Murthy et al., 2018). Bengali, Tamil, and Telugu have seen similar attention, with recent BERT-based systems reaching F1 scores of 75–80 on WikiAnn (Pan et al., 2017) and other silver-standard corpora. The systematic survey by Malarkodi et al. (2016) and the more recent analysis by Bhattacharjee et al. (2019) document these developments in detail.

The most important recent development for Indic NER is the emergence of pretrained encoders specifically designed for Indian languages. MuRIL (Khanuja et al., 2021), pretrained on 17 Indian languages including Assamese with a combination of monolingual, translated, and transliterated data, has become the default backbone for Indic NER and reached the strongest

single-encoder scores across Hindi, Bengali, Tamil, and Telugu at release. IndicBERT (Kakwani et al., 2020) and its successor IndicBERTv2 (Doddapaneni et al., 2023) target the same setting and provide competitive alternatives, with IndicBERTv2 reaching F1 scores in the 80–85 range on the Naamapadam Hindi split. Language-agnostic multilingual encoders (mBERT, XLM-RoBERTa) remain useful reference points but consistently underperform the Indic-specialised encoders on Indic NER benchmarks, which is consistent with the pattern we observe on NerAxom in Section 6.

For Assamese specifically, NER research remains in its early stages. Sharma et al. (2010) contributed the earliest reported work: a manually tagged corpus of approximately 50,000 words together with a set of hand-authored rules for Person, Location, and Organization tagging. In a follow-up study, Sharma et al. (2014) adopted a two-tier strategy that combined a CRF-based statistical layer with rule-based post-processing, reporting F1 scores in the range of 60–65 for the three principal categories. Talukdar et al. (2014) applied CRF-based supervised learning with handcrafted features on a similarly small manually annotated corpus, reporting F1 scores of approximately 65 for Person and Location and lower scores for Organization. Sharma et al. (2016) extended the two-tier approach with hybrid rule-and-machine-learning strategies. A common limitation across this early work is the absence of a public, gold-standard Assamese NER corpus, which prevented direct comparison between systems and slowed adoption of neural methods.

The AsNER corpus (Pathak et al., 2022) addressed part of this gap by releasing a publicly available Assamese NER dataset of approximately 98,000 tokens with five entity categories: Person, Location, Organization, Number, and Miscellaneous. The reported baseline systems on AsNER include BiLSTM-CRF and a fine-tuned XLM-RoBERTa model, with the strongest baseline reaching approximately 74 F1 across the five categories. Naamapadam (Mhaske et al., 2023) additionally provides silver-standard Assamese NER annotations projected from English, but with a limited category inventory and known projection noise. Neither corpus covers Date, Occupation, or Work-of-Art entities, which are frequent in biographical text and which NerAxom introduces as first-class categories. The AsNER category inventory is also narrower (five vs. seven), and its Miscellaneous class merges semantically heterogeneous entities that our design deliberately separates (Section 3.1). NerAxom therefore complements rather than duplicates AsNER: the two corpora have comparable size (102,857 vs. 98,569 tokens), disjoint category coverage in three of seven cases, and different source domains.

2.3 Multi-Granularity Representations for NER

The combination of multiple representational granularities has a well-established track record in NER, particularly for morphologically rich languages where any single granularity leaves diagnostic information on the table. Word-level embeddings encode lexical semantics but cannot represent out-of-vocabulary tokens or morphological variation within a word. Character-level encoders address both limitations by learning representations directly from the character sequence, and were shown by Dos Santos and Guimaraes (2015) to yield consistent NER gains for Portuguese and Spanish. Subword-level tokenisation (Sennrich et al., 2016) occupies an intermediate position and became the dominant granularity with the rise of BERT-family models, but as observed by Clark et al. (2022) and confirmed for Indic languages by Bhattacharjee et al. (2019), subword tokenisers trained on multilingual corpora frequently fragment morphologically meaningful units in ways that a character-level encoder can recover.

The hybrid character-word design of Lample et al. (2016) and Ma and Hovy (2016) was the first widely adopted multi-granularity architecture for NER. It concatenates character-level BiLSTM or CNN features with pretrained word embeddings before passing them to a BiLSTM-CRF tagger. This design has been extended in several directions. Aguilar et al. (2017) proposed a multi-task variant that jointly models character, word, and sentence levels within a shared encoder, reporting F1 gains of 2–3 points on English NER for social media data. Straková et al.

(2019) showed that combining contextual embeddings (Flair, ELMo, BERT) with character-level features yielded consistent improvements across four European languages, reaching 93.38 F1 on CoNLL-2003 English. More recent work has extended the multi-granularity design to span-level features: Bao et al. (2023) integrated BERT representations with span-level features in a unified MRC framework for nested NER, and Li et al. (2020) generalised span-based decoding to flat, nested, and discontinuous NER within a single model.

For Indian languages specifically, character-transformer hybrid designs have been examined in a small number of studies. Murthy et al. (2018) showed that character-level features complement multilingual BERT for Hindi and Bengali NER, and Mhaske et al. (2023) used character-aware inputs in their Naamapadam baselines. To the best of our knowledge, no prior work has applied a character-CNN plus MuRIL hybrid to Assamese NER, and no prior work has combined character-level features with an explicit entity-span attention mechanism for any Indic language. MGF-NER contributes these design decisions in a single system and evaluates them on the publicly available NerAxom corpus introduced in our concurrent work (Sarmah et al., 2026).

2.4 Transfer Learning for Low-Resource NER

Transfer learning is the dominant paradigm for low-resource NER because gold-standard annotation at scale is prohibitively expensive for most languages. Two families of approaches are relevant. Cross-lingual transfer leverages multilingual pretraining to enable zero-shot or few-shot NER in a target language (Wu and Dredze, 2019; Pires et al., 2019; Conneau et al., 2020). Wu and Dredze (2019) showed that multilingual BERT transfers surprisingly effectively across typologically diverse languages, and Pires et al. (2019) analysed the conditions under which cross-lingual transfer is most likely to succeed. For Indian languages, cross-lingual transfer from English or Hindi has yielded mixed results, with transfer quality strongly dependent on linguistic similarity and script relatedness (Bhattacharjee et al., 2019).

Language-adaptive fine-tuning strategies complement cross-lingual transfer by injecting target-language knowledge into an already-pretrained multilingual model. Adapter modules (Pfeiffer et al., 2020) and continued pretraining on target-language corpora (Gururangan et al., 2020) both narrow the gap between a general-purpose multilingual encoder and a target-language specialised one, at a fraction of the cost of full pretraining. For Assamese, continued pretraining is a natural avenue that we do not pursue in this work; the design of MGF-NER instead combines the pretrained MuRIL backbone with lightweight language-specific components (Char-CNN, Entity Span Attention) that add Assamese-specific inductive biases without any additional pretraining.

The distinction between these two families matters for the interpretation of our results. The 43 F1 point gap between BiLSTM CRF without pretraining and MGF NER using pretrained MuRIL with Char CNN, span attention, and a CRF layer in Section 6 is primarily attributable to the transfer of pretrained multilingual representations and only secondarily to the additional Assamese specific inductive bias. The 1.57-F1 gap between MuRIL alone and MGF-NER, by contrast, isolates the contribution of the Assamese-specific components, and it is this gap that the ablation study in Section 6.3 decomposes.

3 NerAxom Dataset

In this paper, we use the NerAxom annotated dataset for Named Entity Recognition (NER) in Assamese. NerAxom was developed specifically for Assamese NER and was first introduced and published by the authors in our concurrent work (Sarmah et al., 2026). The NerAxom dataset is publicly available at **this repository**. We retain the following description of the dataset to provide readers with the necessary background on its construction, annotation scheme, and statistics.

NerAxom consists of seven entity categories: Person, Location, Organization, Date, Work_of_Art, Occupation, and Number. Because biographies in Wikipedia¹ contain information about a wide range of entities, we selected them as a primary source for constructing the NerAxom dataset. This provides a broad and diverse collection of named entities across different professions and domains, making the dataset suitable for evaluating NER models under varied linguistic and semantic contexts. Moreover, the structured nature of Wikipedia biographies facilitates the extraction of information for both training and evaluation.

A significant portion of the dataset comes from biographies written in Assamese on Wikipedia. Additionally, we added person names, location names, and organization names from AsNER (Pathak et al., 2022), which were relabelled according to the BIO tagging scheme. The resulting NerAxom dataset contains 4,178 sentences comprising 102,857 tokens. The dataset includes seven categories of named entities, as detailed in Table 1, along with corresponding examples and definitions for each entity.

Table 1: Entity Names, Definitions, and Examples of NerAxom

Entity Name	Definition	Examples
Person (PER)	Defines the name of a person. Entities could be individual or with a rank or designation. Sometimes honorific addressing terms before a person’s name also act as person entities.	আনন্দ চন্দ্ৰ আগৰৱালা (Anand Chandra Agrawal), ড° দিলীপ কুমাৰ দত্ত (Dr. Dilip Kumar Dutta), ৰসৰাজ লক্ষ্মীনাথ বেজবৰুৱা (Rasraj Laxminath Bezbaruah)
Location (LOC)	Indicates roads, trajectories, regions i.e. villages, towns, cities, countries etc. Also includes natural locations such as mountains, rivers, nature reserves, national parks etc.	গুৱাহাটী (Guwahati), মাজুলী (Majuli), কাজিৰঙা ৰাষ্ট্ৰীয় উদ্যান (Kaziranga National Park), ব্ৰহ্মপুত্ৰ (Brahmaputra)
Organization (ORG)	Defined by companies, enterprises, banks, societies, schools, colleges, universities, newspapers.	কেন্দ্ৰীয় বিদ্যালয় (Kendriya vidyalaya), দৰং মহাবিদ্যালয় (Darrang college), তেজপুৰ বিশ্ববিদ্যালয় (Tezpur university), অসম গ্ৰামীণ বিকাশ বেংক (Assam Gramin Vikash), নিয়মীয়া বাৰ্তা (Niyamiya barta)
Date (DATE)	Entities indicating year, months, date.	১৯২২ চনত (1922), ২৬ ডিচেম্বৰ (26th December), জুলাই (July)
Work-Of-Art (WOA)	Specifies the names or titles of books, songs, films, stories, albums, musicals, TV shows, awards, and similar entities.	জোনাকী ৰাতি (Jonaki Rati), জীয়ে কিউ (Jiye kyu), তেং চাৰ্লে (Tengo Charley), মেৰী কম (Mary Kom)
Occupation (OCC)	Indicates a definite designation	শিক্ষক (Teacher), পৰিদৰ্শক (Inspector), উপ-পৰিদৰ্শক (Sub-Inspector)
Number (NUM)	Defines the quantity of measurement.	৭ বছৰ (7 years), ১৪ কেজি (14kg), ২ জন (2)

3.1 Annotation Scheme

Since Assamese does not have a widely accepted inventory of named entity categories for corpus development, the NerAxom dataset was annotated using guidelines developed by the authors by adapting two established resources: the Groningen Meaning Bank (GMB) (Bos et al., 2017)

1. <https://as.wikipedia.org/wiki>

and the CoNLL-2003 shared-task guidelines (Sang and De Meulder, 2003). The guidelines were refined during a pilot annotation of 300 sentences before being applied to the full corpus.

Entity inventory. We define seven entity categories: Person (PER), Location (LOC), Organization (ORG), Date (DATE), Number (NUM), Occupation (OCC), and Work-of-Art (WOA). Definitions and examples for each category are given in Table 1. Non-entity tokens are labelled 0 and account for approximately 39% of the corpus. In contrast to CoNLL-style datasets, we deliberately avoid a general *Miscellaneous* category. Entities that would ordinarily be placed under *Miscellaneous*, such as creative works, temporal expressions, and quantities, are instead assigned to WOA, DATE, and NUM respectively. This design produces finer-grained supervision without materially increasing annotation cost, and it aligns with the observation that the Miscellaneous class in CoNLL-2003 has historically been the hardest to model reliably because of its semantic heterogeneity.

Rationale for the Work-of-Art category. The Work of Art (WOA) label covers titles of creative works, including books, novels, poems, songs, films, television programmes, and awards. These entities occur frequently in Assamese biographies, accounting for 11.8% of all tokens and 13.7% of annotated entity spans in NerAxom, but are not represented by the standard Person, Location, or Organization categories. Rather than introducing multiple fine-grained labels (e.g., Book, Film, Song, Award), we adopted a single unified WOA category for three reasons. First, at the corpus scale considered here, most fine-grained subclasses would contain only tens or low hundreds of examples, which is too sparse to support reliable supervised learning. Second, a unified creative-works label is consistent with widely adopted annotation frameworks including the GMB, which treats creative works as a single semantic class. Third, a unified label preserves within-category annotation consistency while still providing a discriminative training signal for a semantically coherent group of entities. We regard a further split of WOA into finer subclasses as a natural extension for future work as the corpus grows.

Tagging scheme. Several tagging schemes have been proposed for NER, including IO, BIO, BIOES, BILUO, IOE, IOE2, and BIO2 (Alshammari and Alanazi, 2021). Following common practice for post-positional and morphologically rich languages such as Hindi, Bengali, and Japanese (Haque et al., 2023; Malik and Sarwar, 2016). Under BIO, the first token of an entity mention receives a *B-type* tag, subsequent tokens within the same mention receive *I-type* tags, and tokens outside any entity receive the tag 0. The scheme treats each entity as a contiguous span with an explicit beginning, which is a natural fit for the multi-word structure of Assamese organisation names, work titles, and person names with honorifics.

Illustrative example. Consider the Assamese sentence তেজপুৰ বিশ্ববিদ্যালয় ১৯৯৪ চনত প্রতিষ্ঠা কৰা হৈছিল (“Tezpur University was established in 1994”). Under the BIO scheme, this sentence is annotated token by token as shown in Table 2. The entity তেজপুৰ বিশ্ববিদ্যালয় (*Tezpur University*) forms a two-token ORG span, and ১৯৯৪ চনত (*in 1994*) forms a two-token DATE span in which the year is inflected with the locative suffix -ত. The example illustrates two properties of Assamese that are directly relevant to NER: multi-word institutional names are common, and post-positional suffixation is applied to date and numeric expressions, which the model must handle without breaking the entity boundary.

Annotation process and annotators. The corpus was annotated by two senior-undergraduate annotators who are native speakers of Assamese. Both annotators had prior coursework in natural language processing and computational linguistics, both received a written copy of the annotation guidelines, and both attended a one-day training session on a shared set of 50 warm-up sentences before beginning the main task. Annotation then proceeded in two phases. In the first phase, the two annotators independently labelled the entire corpus using the BIO scheme, without access to each other’s decisions. In the second phase, all sentences on which the two annotators disagreed were reviewed jointly with a linguist consultant who is

Table 2: Illustration of BIO tagging on an Assamese sentence.

Token	Transliteration	BIO Tag
তেজপুৰ	Tezpur	B-ORG
বিশ্ববিদ্যালয়	Bishwabidyalay	I-ORG
১৯৯৪	1994	B-DATE
চনত	chonot	I-DATE
প্ৰতিষ্ঠা	pratistha	O
কৰা	kora	O
হৈছিল	hoisil	O

a native Assamese speaker and a domain expert in Assamese grammar. Disagreements were resolved by discussion, with the linguist acting as the final arbiter; the outcomes of these discussions were recorded in a written appendix to the guidelines so that similar cases would be handled consistently thereafter. The adjudicated annotations from this second phase constitute the gold-standard version of the corpus. Cohen’s κ over the two annotators’ independent entity-type labels was 0.87, indicating substantial agreement; the corresponding boundary-level statistics and a discussion of what can and cannot be recomputed from the retained materials are given in Section 3.2.

3.2 Dataset Statistics

The entity spans of NerAxom distributed across the seven entity categories defined in Table 1. Table 3 reports the token-level distribution under the BIO tagging scheme. Under BIO, each entity mention begins with a single B- tag and continues with zero or more I- tags. Consequently, the number of B- tags for a given category corresponds directly to the number of annotated entity spans of that type, while the total (B+I) reflects the number of tokens participating in entities of that type. Tokens that do not belong to any entity are labelled O.

Table 3: Token-level BIO tag distribution in the NerAxom corpus.

NE Tag	B-Tag	I-Tag	Total
PER	16,827	7,940	24,767
LOC	7,465	3,871	11,336
ORG	1,956	3,516	5,472
NUM	717	5	722
DATE	4,480	2,774	7,254
WOA	5,072	7,018	12,090
OCC	399	752	1,151
O	—	—	40,065
Total	36,916	25,876	102,857

Of the 102,857 tokens, 40,065 tokens representing 38.95% are non-entity tokens, while 62,792 tokens representing 61.05% belong to entity spans. Non-entity tokens are not vacuous: they provide the local context on which BIO decisions and entity-type disambiguation depend. Across the seven entity categories, the distribution of entity mentions is naturally imbalanced. Person (PER) accounts for 45.6% of all entity spans, followed by Location (LOC) at 20.2%, Work-of-Art (WOA) at 13.7%, Date at 12.1%, Organization (ORG) at 5.3%, Number (NUM) at 1.9%, and Occupation (OCC) at 1.1%. The two rare categories, OCC and NUM, present the greatest

learning challenge and are analysed separately in the per entity results reported in Table 9 of Section 6.

Because a single entity mention can span multiple tokens, token-level counts and entity-level counts differ substantially and should not be conflated. Table 4 makes the distinction explicit and reports the average span length per category. Person names average 1.47 tokens per mention, consistent with a mix of single-name and multi-token full-name references. Organization entities average 2.80 tokens per mention, reflecting the multi-word structure of institutional names in Assamese. Work-of-Art (2.38 tokens per mention) and Occupation (2.88 tokens per mention) are similarly multi-token, since creative-work titles and descriptive occupational designations typically span several words. Location entities have an average length of 1.52 tokens, Date entities average 1.62 tokens, and Number entities are almost exclusively single token with an average length of 1.01 tokens.

Table 4: Entity-level statistics: number of annotated spans, share of total spans, and average span length in tokens.

Category	Entity Spans	Share of Total	Avg. Length (tokens)
PER	16,827	45.6%	1.47
LOC	7,465	20.2%	1.52
WOA	5,072	13.7%	2.38
DATE	4,480	12.1%	1.62
ORG	1,956	5.3%	2.80
NUM	717	1.9%	1.01
OCC	399	1.1%	2.88
Total	36,916	100.0%	1.70

We partition the corpus into training, validation, and test sets in an 80/10/10 ratio at the sentence level, resulting in 3,342 training, 418 validation, and 418 test sentences. Sentence-level partitioning preserves intra-sentence coherence and yields entity distributions that are approximately proportional across the three splits, ensuring that all seven categories are represented in every split. Table 5 reports the per-category entity-span counts for each split. Even for the rarest category (OCC), the test split retains 40 entity mentions, which we take into account when interpreting per-category F1 scores in Section 6.

Table 5: Per-category entity-span counts across training, validation, and test splits (80/10/10 sentence-level partition).

Category	Train	Validation	Test	Total
PER	13,462	1,682	1,683	16,827
LOC	5,972	746	747	7,465
WOA	4,058	507	507	5,072
DATE	3,584	448	448	4,480
ORG	1,565	195	196	1,956
NUM	574	71	72	717
OCC	319	40	40	399
Total	29,534	3,689	3,693	36,916

4 Proposed Methodology

This section presents the proposed Multi-Granularity Fusion Named Entity Recognition (MGF-NER) framework for Assamese. We first formulate the NER task as a sequence labeling problem and then describe each component of the architecture, including the character-level representation, transformer-based contextual encoder, multi-granularity fusion mechanism, entity span attention module, and CRF decoder. Finally, we outline the training objective and inference procedure used to optimize the model.

4.1 Problem Formulation

Let $S = \{w_1, w_2, \dots, w_n\}$ denote an input sentence consisting of n words. Named entity recognition on S is formulated as a token-level sequence labelling problem: the goal is to learn a mapping

$$f : S \rightarrow Y,$$

where $Y = \{y_1, y_2, \dots, y_n\}$ is the corresponding label sequence. Each label y_i belongs to a predefined tag set L under the BIO (Beginning-Inside-Outside) annotation scheme:

$$L = \{O\} \cup \{B-t, I-t \mid t \in T\},$$

with $T = \{\text{PER, LOC, ORG, DATE, NUM, OCC, WOA}\}$ the seven entity types defined in Table 1. This formulation treats every token as a decision point and every entity mention as a contiguous span with an explicit beginning, which matches the structure of NerAxom as described in Section 3.2.

4.2 Architecture Overview and Design Rationale

MGF-NER is a hierarchical pipeline that combines six components in the following order: a (1) Character-level Convolutional Neural Network, a (2) Transformer-based Contextual Encoder (MuRIL), a (3) Multi-Granularity Fusion Layer, an (4) Entity Span Attention Module, a (5) Bidirectional LSTM Encoder, and a (6) Conditional Random Field Decoder. Figure 1 illustrates the complete data flow.

The design is guided by a single principle: named entity recognition for a morphologically rich, low-resource language cannot rely on any single representational granularity in isolation. Subword-based multilingual transformers such as MuRIL provide strong contextual semantics but tokenise Assamese with a WordPiece vocabulary that does not respect morpheme boundaries, so morpheme-level signals for category discrimination (case markers, honorifics, tense-aspect suffixes) are attenuated. Character-level encoders recover those signals but are blind to sentence-level context. Token-level architectures are efficient but do not explicitly model the multi-word span structure that is common in Assamese organisation names and creative-work titles. And per-token classification heads decouple the label decisions across positions, which permits BIO-invalid sequences (for example, an $I-t$ tag directly following an O tag) that a discriminative structured decoder would rule out by construction.

Each component in the pipeline addresses one of these gaps. The Char-CNN carries morphological information that survives independent of WordPiece fragmentation; MuRIL contributes cross-lingually pretrained contextual semantics; the fusion layer resolves the two representations into a single per-word vector; the Entity Span Attention module makes multi-word span boundaries explicit; the BiLSTM smooths the resulting features across positions; and the CRF layer enforces valid BIO transitions. The remainder of Section 4 formalises each component and states, for each one, the specific role it plays and why alternatives are less well suited. The ablation results in Section 6.3 then verify empirically that the components are complementary rather than redundant.

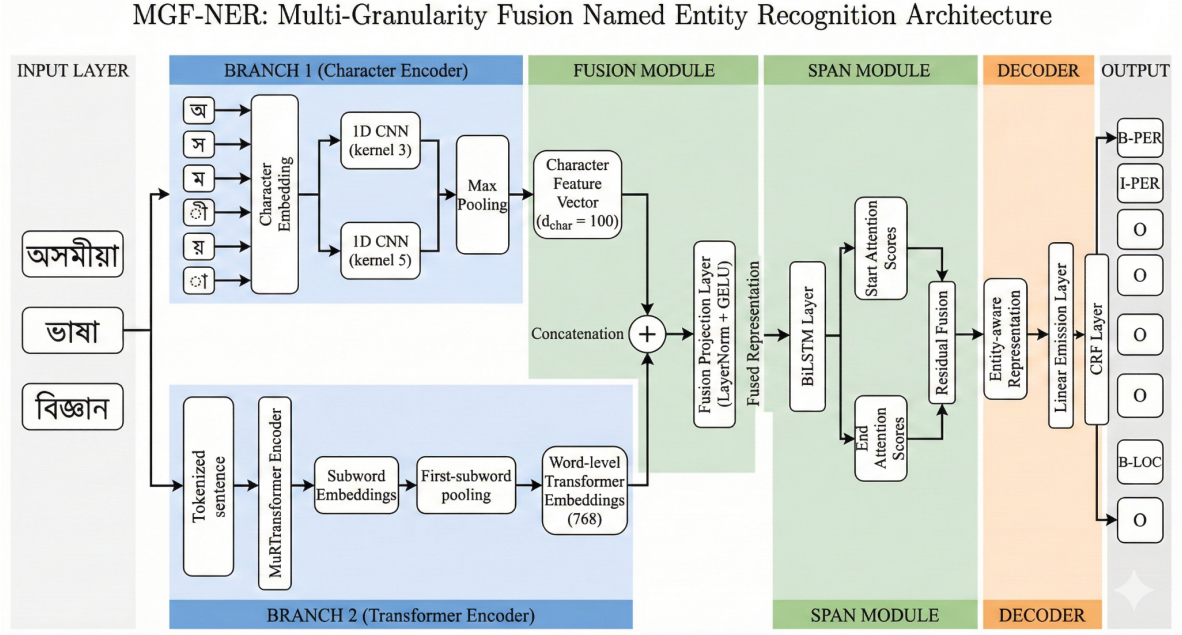


Figure 1: Overall architecture of MGF-NER. Character sequences of each word are encoded by the Char-CNN into $\mathbf{x}_i^{char}$. The full sentence is passed through MuRIL and the first-subword representation of each word is used as $\mathbf{x}_i^{trans}$. The two representations are concatenated, projected, and normalised into the fused representation $\mathbf{h}_i^{fused}$. The Entity Span Attention module refines these features with boundary-aware attention, and the CRF layer decodes a globally consistent BIO label sequence.

4.3 Character-Level Representation

Assamese is a morphologically rich, agglutinative Indo-Aryan language in which productive suffixation carries substantial entity-discriminative information. Case markers such as -ত (locative) and -ৰ (genitive) attach to place names; honorific and vocational suffixes discriminate Person from Occupation entities; and multi-morphemic compound formations are pervasive in institutional names and creative-work titles. Preserving these sub-word signals requires a representation that operates strictly below the subword granularity used by the transformer backbone.

Choice of character-level encoding over alternative subword schemes. A reasonable alternative to a character-level encoder is a second subword tokenisation stream, such as an independent byte pair encoding (BPE) or SentencePiece model trained on Assamese and concatenated with MuRIL’s representation. We considered this option and adopted a character-level Convolutional Neural Network (Char-CNN) instead for three reasons. First, MuRIL’s WordPiece tokeniser already provides a subword representation for the sentence, so an additional BPE stream would add information at the same granularity rather than a complementary one; the marginal benefit is expected to be small. Second, multilingual WordPiece tokenisers, including MuRIL’s, do not respect Assamese morpheme boundaries: inflected forms and compounds are frequently fragmented in ways that split across meaningful morphemes, as illustrated for representative Assamese words in Table 6. A character-level encoder operates strictly below the WordPiece level and can therefore recover morpheme-consistent signals regardless of how a given word happened to be fragmented at the subword level. Third, character level CNNs are parameter efficient, requiring fewer than 100K parameters in our configuration, and do not require additional pretraining, whereas a dedicated Assamese BPE encoder would require its own pretraining corpus and would either replace or duplicate the MuRIL backbone. The Char-CNN

is therefore both linguistically justified for Assamese and complementary to MuRIL’s subword stream by construction. This design follows the character-word hybrid strategy first shown to be effective for morphologically rich languages by Lample et al. (2016) and Ma and Hovy (2016), and used in the NER context specifically by Dos Santos and Guimaraes (2015) and Aguilar et al. (2017).

Table 6: Illustrative WordPiece fragmentation of Assamese words by MuRIL’s tokeniser. Fragments that split across morpheme boundaries are marked with a leading ‘##’. The Char-CNN encoder is not affected by this fragmentation and processes each word as a full character sequence.

Assamese word	Gloss	MuRIL WordPiece segmentation
বিশ্ববিদ্যালয়	university	বিশ্ব ##বিদ্যালয়
মহাবিদ্যালয়	college	মহা ##বিদ্যালয়
ৰাষ্ট্ৰীয়	national	ৰাষ্ট্ৰ ##ীয়
কাজিৰঙাত	at Kaziranga (loc.)	কাজি ##ৰঙা ##ত

Formalisation. For each word w_i , let

$$\mathbf{c}_i = \{c_{i,1}, c_{i,2}, \dots, c_{i,m}\}$$

denote its constituent character sequence of length m . Each character $c_{i,j}$ is mapped to a dense vector using a character embedding matrix

$$\mathbf{E}^{char} \in \mathbb{R}^{|\mathcal{V}_{char}| \times d_c},$$

where $|\mathcal{V}_{char}|$ is the size of the character vocabulary (approximately 100 characters, including Assamese Unicode block symbols, diacritics, digits, and punctuation) and $d_c = 30$ is the character embedding dimension. The embedding for character $c_{i,j}$ is

$$\mathbf{e}_{i,j}^{char} = \mathbf{E}^{char}[c_{i,j}] \in \mathbb{R}^{d_c}.$$

The dimension d_c is a learned embedding size and is chosen independently of $|\mathcal{V}_{char}|$; because the mapping is a learned lookup rather than a one-hot encoding, d_c need not equal (and is not required to exceed) the character vocabulary size. A 30-dimensional embedding was selected empirically as a balance between expressiveness and parameter economy, matching the setting used in prior character-CNN work for NER (Ma and Hovy, 2016).

To capture morphological patterns at multiple granularities, the embedded character sequence is processed in parallel by convolutional filters with kernel sizes $\mathcal{K} = \{3, 5\}$. For a filter with kernel size $k \in \mathcal{K}$,

$$\mathbf{h}_{i,j}^{(k)} = \text{ReLU}\left(\mathbf{W}^{(k)} * \mathbf{E}_i^{char}[j : j + k - 1] + \mathbf{b}^{(k)}\right),$$

where $\mathbf{W}^{(k)} \in \mathbb{R}^{d_f \times k \times d_c}$ and $\mathbf{b}^{(k)} \in \mathbb{R}^{d_f}$ are learnable convolutional parameters and $d_f = 50$ is the number of filters per kernel size. Max-over-time pooling selects the most salient feature per filter,

$$\hat{\mathbf{h}}_i^{(k)} = \max_j \mathbf{h}_{i,j}^{(k)} \in \mathbb{R}^{d_f},$$

and the final character-level representation of word w_i is obtained by concatenating the pooled outputs of the two kernel sizes:

$$\mathbf{x}_i^{char} = \left[\hat{\mathbf{h}}_i^{(3)} \oplus \hat{\mathbf{h}}_i^{(5)}\right] \in \mathbb{R}^{d_{char}},$$

where $d_{char} = |\mathcal{K}| \times d_f = 100$. The value 100 in the concatenated output should not be confused with the character embedding dimension $d_c = 30$: the former is a property of the CNN output (two kernel sizes with 50 filters each), while the latter is a property of the character lookup table.

4.4 Transformer-Based Contextual Encoding

For contextual semantics, we use MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), a BERT-base transformer pretrained on 17 Indian languages including Assamese. MuRIL is preferred over language-agnostic multilingual encoders (mBERT, XLM-RoBERTa) for two reasons. First, MuRIL’s pretraining corpus includes native Assamese Wikipedia and Common Crawl text together with translated and transliterated content, whereas mBERT and XLM-R include Assamese only as a minor constituent of a much larger multilingual mix. Second, MuRIL’s tokeniser is trained jointly on the Indian language corpus and yields more meaningful subword pieces for Assamese than the language-agnostic tokenisers, though it still fragments long compound forms as shown in Table 6. The empirical comparison in Section 6 confirms that MuRIL is a stronger single-encoder baseline than either mBERT or XLM-R on NerAxom.

Given an input sentence $S = \{w_1, w_2, \dots, w_n\}$, MuRIL’s WordPiece tokeniser produces a subword sequence

$$S^{sub} = \text{Tokenize}(S) = \{s_1, s_2, \dots, s_T\},$$

with T subword tokens in total ($T \geq n$). MuRIL encodes S^{sub} using stacked transformer layers whose self-attention operation is defined per head as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

with query, key, and value matrices Q, K, V and key dimensionality d_k . The multi-head outputs are concatenated and passed through a position-wise feed-forward network at every layer, yielding contextualised subword representations

$$\mathbf{H}^{trans} = \text{MuRIL}(S^{sub}) \in \mathbb{R}^{T \times d_{model}},$$

with hidden dimension $d_{model} = 768$.

4.5 Subword-to-Word Alignment

Because MuRIL operates on subword pieces while our label sequence is defined at the word level, we align subword representations back to word-level representations. For each word w_i , we take the contextual representation of the first subword token belonging to w_i :

$$\mathbf{x}_i^{trans} = \mathbf{H}^{trans}[\alpha(i)] \in \mathbb{R}^{d_{model}},$$

where $\alpha(i)$ is the index of the first subword in S^{sub} that originates from word w_i . First-subword alignment is preferred over pooling strategies (mean or max over all subwords of a word) because it follows the standard token-label alignment convention adopted in BERT-based token-level classification, where the representation of the first sub-token is used as the input to the token-level classifier Devlin et al. (2019). This approach also provides a deterministic word-level anchor point regardless of how many subwords a word is fragmented into.

4.6 Multi-Granularity Fusion

The character-level and transformer-based representations are combined into a single per-word vector through a two-step fusion. In the first step, we concatenate the two representations:

$$\mathbf{z}_i = \left[\mathbf{x}_i^{trans} \oplus \mathbf{x}_i^{char} \right] \in \mathbb{R}^{d_{model} + d_{char}}. \quad (1)$$

The vector $\mathbf{z}_i$ in Equation (1) juxtaposes the two representations without mixing them; its dimensionality of $d_{model} + d_{char} = 868$ is inconvenient for downstream layers, and its two halves reside in different learned subspaces.

In the second step, $\mathbf{z}_i$ is projected into a shared d_{fuse} -dimensional feature space, layer-normalised, and passed through a GELU non-linearity:

$$\mathbf{h}_i^{fused} = \text{GELU}\left(\text{LayerNorm}(\mathbf{W}_{fuse} \mathbf{z}_i + \mathbf{b}_{fuse})\right), \quad (2)$$

where $\mathbf{W}_{fuse} \in \mathbb{R}^{d_{fuse} \times (d_{model} + d_{char})}$ is a learnable projection matrix, $\mathbf{b}_{fuse} \in \mathbb{R}^{d_{fuse}}$ is a learnable bias vector, and $d_{fuse} = 384$ is the fusion projection dimension listed in Table 7. The vector $\mathbf{z}_i$ in Equation (2) is the same concatenated vector defined in Equation (1); the two-step formulation (concatenate, then project) is used to keep the two operations distinct in the discussion. The notation used in this subsection separates *vector* labels (superscripts, such as $\mathbf{x}_i^{trans}$, $\mathbf{x}_i^{char}$, $\mathbf{h}_i^{fused}$) from *parameter* labels (subscripts, such as $\mathbf{W}_{fuse}$, $\mathbf{b}_{fuse}$); the two categories refer to different objects (activations vs. learnable parameters), which is why $\mathbf{h}^{fused}$ is written with a superscript while $\mathbf{W}_{fuse}$ and $\mathbf{b}_{fuse}$ are written with a subscript. The learnable parameters $\mathbf{W}_{fuse}$ and $\mathbf{b}_{fuse}$ do not appear in the hyperparameter table because they are learned end-to-end during training rather than set by the practitioner; only the projection dimension d_{fuse} is a hyperparameter.

The role of the projection step is twofold: it aligns the two representational halves into a common vector space where their information can be jointly attended to by downstream layers, and it reduces the dimensionality from 868 to 384, which reduces the parameter count of every subsequent module by roughly a factor of two.

4.7 Entity Span Attention Module

Named entities in Assamese are frequently multi-token, particularly for ORG and WOA categories (Section 3.2). Token-level classification alone does not explicitly encode boundary decisions, and the resulting per-token predictions can produce entities with truncated or over-extended spans. The Entity Span Attention module addresses this gap by injecting an explicit boundary-aware signal into the per-word representation.

The fused word representations are first passed through a bidirectional LSTM to capture short-range sequential dependencies:

$$\mathbf{h}_i^{lstm} = \text{BiLSTM}(\mathbf{h}_i^{fused}).$$

Two learnable attention vectors, $\mathbf{w}_{start}$ and $\mathbf{w}_{end}$, produce boundary-aware attention scores at every position:

$$\alpha_i^{start} = \frac{\exp(\mathbf{w}_{start}^\top \mathbf{h}_i^{lstm})}{\sum_{j=1}^n \exp(\mathbf{w}_{start}^\top \mathbf{h}_j^{lstm})}, \quad \alpha_i^{end} = \frac{\exp(\mathbf{w}_{end}^\top \mathbf{h}_i^{lstm})}{\sum_{j=1}^n \exp(\mathbf{w}_{end}^\top \mathbf{h}_j^{lstm})}.$$

The scalars α_i^{start} and α_i^{end} are trained implicitly to peak at positions that are, respectively, entity start and entity end candidates; no explicit boundary supervision is added to the loss, but the CRF layer’s use of these signals during BIO decoding provides an indirect signal during training. The span-aware representation is obtained by concatenating the LSTM output with the two boundary scores, projecting, and adding a residual connection with layer normalisation:

$$\mathbf{h}_i^{entity} = \text{LayerNorm}\left(\mathbf{h}_i^{fused} + \text{Dropout}(\mathbf{W}_{span} [\mathbf{h}_i^{lstm} \oplus \alpha_i^{start} \oplus \alpha_i^{end}])\right).$$

The residual connection from $\mathbf{h}_i^{fused}$ preserves the fused representation as a strong signal even when the boundary attention is uninformative (for example, on single-token entities), so the module contributes additively rather than displacing the underlying representation. The ablation in Section 6.3 confirms this design: removing the module produces the largest per-category degradation on ORG and WOA, which are the categories with the highest average span length.

4.8 CRF-Based Sequence Decoding

A softmax classifier applied independently at every position does not enforce the structural constraints of the BIO scheme. It permits sequences that are invalid by construction, such as an I- t tag following an O tag or an I- t tag following a B- t' tag for $t \neq t'$, both of which correspond to no legitimate entity. A linear-chain Conditional Random Field decoder resolves this by modelling label dependencies explicitly and choosing the highest-scoring globally consistent label sequence.

Given the sequence of span-aware features

$$\mathbf{H}^{entity} = \{\mathbf{h}_1^{entity}, \dots, \mathbf{h}_n^{entity}\},$$

the CRF defines a conditional distribution over label sequences $y = \{y_1, \dots, y_n\}$ as

$$P(y \mid \mathbf{e}) = \frac{\exp(\text{Score}(\mathbf{e}, y))}{\sum_{y' \in \mathcal{Y}_{all}} \exp(\text{Score}(\mathbf{e}, y'))},$$

where $\mathbf{e}$ denotes the emission scores produced from $\mathbf{H}^{entity}$ through a linear projection into the tag space, $\mathcal{Y}_{all}$ is the set of all valid label sequences, and

$$\text{Score}(\mathbf{e}, y) = \sum_{i=1}^n \mathbf{e}_i[y_i] + \sum_{i=1}^{n-1} \mathbf{A}[y_i, y_{i+1}] + \mathbf{A}_{start}[y_1] + \mathbf{A}_{end}[y_n].$$

The learnable transition matrix $\mathbf{A}$ and the start/end transition vectors $\mathbf{A}_{start}, \mathbf{A}_{end}$ are trained jointly with the rest of the network. BIO validity is enforced by assigning prohibitively low transition scores to structurally invalid transitions before decoding, so the Viterbi search at inference time cannot select any sequence outside $\mathcal{Y}_{all}$.

The CRF is the second most important component of MGF-NER in the ablation study, and its per-category impact concentrates on the two categories with the largest average span length (ORG and WOA), which is consistent with its role as a structural constraint rather than a representational one.

4.9 Training and Inference

Training objective. The model is trained end-to-end by minimising the negative log-likelihood of the gold label sequence under the CRF,

$$\mathcal{L}_{NLL} = -\log P(y \mid \mathbf{e}),$$

computed with the standard forward algorithm. All components (Char-CNN embeddings and convolutions, MuRIL parameters, fusion projection, span-attention parameters, BiLSTM, and CRF transitions) receive gradient updates. The two learning rates specified in Section 5.2 are applied group-wise so that the pretrained MuRIL parameters and the from-scratch modules can converge at appropriate rates.

Inference. At test time, the most probable label sequence is recovered by Viterbi decoding,

$$\hat{y} = \arg \max_{y \in \mathcal{Y}_{all}} \text{Score}(\mathbf{e}, y).$$

The restriction to $\mathcal{Y}_{all}$, together with the CRF transition matrix, guarantees that every predicted sequence is a valid BIO tagging with no structurally impossible transitions.

5 Experimental Setup

This section describes the experimental configuration used to evaluate the proposed MGF-NER framework. We introduce the baseline models selected for comparison, present the implementation details and training hyperparameters, and describe the evaluation protocol adopted throughout the experiments. To ensure fair and reproducible comparisons, all models were trained under identical data splits and optimization settings, with performance reported as the mean and standard deviation over five random seeds.

5.1 Baseline Models

We compare MGF-NER against nine baseline systems that span traditional feature-based sequence labelling, deep recurrent architectures, and transformer-based contextual encoders. The baselines are selected to isolate the contribution of different modelling choices: handcrafted features versus learned representations, recurrent versus attention-based encoders, and language-agnostic versus Indic-specialised pretraining. All baselines are evaluated on the same NerAxom train/validation/test splits reported in Table 5 and under the strict entity-level evaluation described in Section 5.3.

Traditional feature-based baseline. *CRF.* A linear-chain Conditional Random Field trained with handcrafted features, including word-shape features, character-level prefixes and suffixes of length up to four, an indicator of whether the token contains Assamese numerals, part-of-speech style contextual features derived from a ± 2 token window, and gazetteer lookups for common Assamese person titles and place-name suffixes. The CRF is trained with L-BFGS optimisation and L2 regularisation. This baseline represents the pre-neural sequence-labelling paradigm and provides a lower bound for what can be achieved without learned representations.

Recurrent deep-learning baseline. *BiLSTM-CRF.* A bidirectional Long Short-Term Memory network with a linear-chain CRF decoder, following the architecture of Huang et al. (2015). Word representations are 300-dimensional embeddings initialised randomly and trained jointly with the model; no pretrained word or subword embeddings (FastText, GloVe, or contextual encoders) are used. The BiLSTM uses a hidden size of 256 per direction and a dropout rate of 0.5. The character-level CNN encoder is not included in this baseline, so that the effect of pretrained representations can be isolated from the sequence-labelling architecture. We include this configuration deliberately as a minimal deep-learning baseline: it measures what a recurrent tagger can learn from the NerAxom training data alone, without transferred knowledge from any external corpus. The reader should therefore interpret the gap between this baseline and the transformer models primarily as evidence of the value of pretrained multilingual representations rather than as a comparison between recurrent and attention-based architectures.

Transformer-based baselines. Seven transformer encoders are fine-tuned on the NerAxom training set using a standard token-classification head with a linear projection into the BIO tag space. All transformer baselines are fine-tuned end-to-end using the AdamW optimiser and the shared hyperparameter configuration described in Section 5.2.

- **mBERT** (Devlin et al., 2019): multilingual BERT (**bert-base-multilingual-cased**), pretrained on Wikipedia in 104 languages, including Assamese, using masked language modelling and next-sentence prediction.
- **XLM-RoBERTa-base** (Conneau et al., 2020): the base variant of Cross-lingual RoBERTa, pretrained on 2.5 TB of filtered CommonCrawl data covering 100 languages with a masked language modelling objective.
- **XLM-RoBERTa-large** (Conneau et al., 2020): the large variant of the same model, with increased depth and hidden size, retained as a stronger cross-lingual reference point.

- **DistilBERT-multi** (Sanh et al., 2019): a distilled version of multilingual BERT designed for computational efficiency, retaining most of mBERT’s cross-lingual transfer capability at roughly 40% of the parameter count.
- **IndicBERTv2** (Doddapaneni et al., 2023): a BERT-style model pretrained on IndicCorp v2, spanning 24 Indian languages, and specifically designed to support downstream tasks in low-resource Indic settings.
- **Bangla-BERT** (Bhattacharjee et al., 2022): a BERT-style model pretrained on Bangla text. Because Assamese and Bangla share a closely related script and considerable orthographic overlap, this baseline directly tests whether script similarity alone yields effective transfer to Assamese, independently of language-specific pretraining data.
- **MuRIL** (Khanuja et al., 2021): the strongest single-encoder baseline, pretrained on 17 Indian languages including Assamese, with a training corpus that combines monolingual, translated, and transliterated data. MuRIL also serves as the transformer backbone of MGF-NER, so that the ablation between MuRIL alone and the full MGF-NER model isolates the contribution of the multi-granularity components introduced in this work.

Seeded evaluation protocol. To ensure that the reported comparisons reflect model behaviour rather than the outcome of a single fortunate initialisation, each of the nine baselines and MGF-NER are trained and evaluated under five random seeds: 42, 123, 3407, 2025, and 9999. For every model, we report the mean and standard deviation of precision, recall, and micro-averaged F1 in Table 8. All models share the same seeded protocol, the same train/validation/test partition, and the same early-stopping criterion based on validation F1. This design allows the differences reported in Section 6 to be interpreted as robust to initialisation variability rather than as artefacts of individual runs.

5.2 Implementation Details

All models are implemented in PyTorch 2.1 using the HuggingFace Transformers library (version 4.36) for transformer baselines, and `pytorch-crf` for CRF decoding. Traditional feature-based CRF baselines are implemented with `sklearn-crfsuite`. Experiments are conducted on a single NVIDIA workstation-class GPU. To ensure reproducibility, random seeds are fixed across Python, NumPy, and PyTorch (including CUDA), and deterministic behaviour is enabled for cuDNN operations.

Optimisation. All neural models are trained with the AdamW optimiser (Loshchilov and Hutter, 2017). For MGF-NER and the transformer baselines, we use separate learning rates for the pretrained transformer parameters and the newly initialised task-specific parameters. Transformer parameters are updated with a peak learning rate of 3×10^{-5} , following common practice for BERT-family fine-tuning. The Character-CNN encoder, fusion projection, entity span attention module, BiLSTM encoder, and CRF layer are updated with a higher peak learning rate of 1.5×10^{-4} , since these modules are trained from scratch and benefit from faster convergence. A linear warmup over the first 10% of training steps is applied, followed by linear decay to zero. Gradients are clipped to a maximum ℓ_2 norm of 1.0.

Training schedule. All models are trained for up to 40 epochs with a batch size of 128 tokens per micro-batch, with early stopping based on validation micro-averaged F1 with a patience of 7 epochs. The best checkpoint on the validation set is retained for test evaluation. Table 7 summarises the shared hyperparameter configuration.

Table 7: Hyperparameter configuration shared across MGF-NER and the transformer-based baselines. Baseline-specific settings (BiLSTM-CRF hidden size, CRF regularisation strength) are documented in Section 5.1.

Hyperparameter	Value
Transformer learning rate	3×10^{-5}
Non-transformer learning rate	1.5×10^{-4}
Batch size	128
Maximum epochs	40
Early stopping patience	7
Warmup ratio	0.1
Weight decay	0.01
Gradient clipping (ℓ_2)	1.0
Dropout rate	0.2
Character embedding dimension	30
Character-CNN kernel sizes	{3, 5}
Character-CNN filters per kernel	50
Fusion projection dimension	384
BiLSTM hidden size per direction	192
Maximum sequence length	128 subword tokens

Seeded evaluation protocol. As stated in Section 5.1, every model (MGF-NER and the nine baselines) is trained and evaluated under five random seeds: 42, 123, 3407, 2025, and 9999. All other components of the pipeline are held fixed across seeds, including the train/validation/test partition, the tokeniser, the hyperparameter configuration, the early-stopping criterion, and the evaluation script. For each seed we record precision, recall, and micro-averaged F1 on the test set at the best-validation checkpoint; the values reported in Table 8 are the mean and standard deviation over the five seeds. This protocol is applied uniformly to all systems in the main results, per-entity results, ablation, and dropout sensitivity analyses, so that variance figures are directly comparable across settings.

Reproducibility. Code, model checkpoints, and the evaluation harness will be released together with the dataset upon acceptance. The release includes the exact random seeds, the version-pinned dependency list, the training and evaluation logs for all reported runs, and the per-seed test-set predictions.

5.3 Evaluation Metrics

We report micro-averaged Precision (P), Recall (R), and F1-score at the entity level under a strict evaluation setting. A predicted entity is counted as correct only if both its boundary (the exact start and end token positions) and its entity type exactly match the corresponding gold annotation; partial matches contribute neither true positives nor partial credit. Micro-averaging weights each entity mention equally, so categories with more mentions contribute proportionally more to the aggregate score, which is standard for NER evaluation and enables direct comparison with prior work on Assamese and other Indic NER benchmarks. Per-category F1, reported in Section 6, complements the micro-averaged view by exposing performance on rare categories such as OCC and NUM that are otherwise dominated in the micro-average by more frequent classes.

Significance testing. To assess whether the observed differences between MGF-NER and the strongest baseline are statistically meaningful, we complement the mean-and-standard-deviation reporting with a paired significance test computed at the sentence level. For each test sentence,

we compute a sentence-level F1 for each model using its best-seed checkpoint. We then apply a paired Wilcoxon signed-rank test (Wilcoxon, 1992) on the paired sentence-level F1 scores; sentence-level pairing is consistent with our entity-level metric because entity mentions are nested within sentences, whereas token-level pairing would violate the independence assumption of standard tests since adjacent tokens within the same entity are highly correlated. Sentence-level testing therefore provides a methodologically sounder assessment of the improvement over MuRIL than a token-level comparison would.

6 Results and Analysis

This section presents the experimental evaluation of the proposed MGF-NER framework on the NerAxom dataset. We first compare MGF-NER with a diverse set of traditional, neural, and transformer-based baselines, followed by analyses of per-entity performance, ablation studies, the influence of model hyperparameters, computational efficiency, and error characteristics. Unless otherwise stated, all reported results correspond to the mean performance over five random seeds to ensure the robustness and reproducibility of the experimental findings.

6.1 Main Results

Table 8 presents a comprehensive comparison of MGF-NER against all nine baselines on the NerAxom test set. Each model is trained and evaluated under five random seeds (42, 123, 3407, 2025, 9999) using the shared protocol described in Section 5.2; we report the mean and standard deviation of precision, recall, and micro-averaged F1 across the five seeds. The Δ F1 column reports the mean F1 difference relative to MGF-NER.

Table 8: Main results on the NerAxom test set. Values are mean $\pm$ standard deviation across five random seeds. All models share the same train/validation/test partition, hyperparameter configuration, and evaluation protocol.

Model	Precision	Recall	F1	Δ F1
MGF-NER (Ours)	84.42 $\pm$ 0.61	86.51 $\pm$ 0.74	85.45 $\pm$ 0.46	—
MuRIL	84.55 $\pm$ 0.87	83.22 $\pm$ 0.96	83.88 $\pm$ 0.79	−1.57
IndicBERTv2	80.61 $\pm$ 1.08	83.65 $\pm$ 1.17	82.10 $\pm$ 0.93	−3.35
XLM-RoBERTa-large	80.31 $\pm$ 1.26	80.57 $\pm$ 1.41	80.44 $\pm$ 1.12	−5.01
XLM-RoBERTa-base	75.21 $\pm$ 1.54	76.03 $\pm$ 1.63	75.62 $\pm$ 1.36	−9.83
mBERT	72.94 $\pm$ 1.81	73.11 $\pm$ 1.74	73.02 $\pm$ 1.48	−12.43
DistilBERT-multi	63.91 $\pm$ 2.03	67.54 $\pm$ 2.18	65.67 $\pm$ 1.84	−19.78
CRF	65.58 $\pm$ 1.72	64.71 $\pm$ 1.95	65.14 $\pm$ 1.57	−20.31
Bangla-BERT	56.18 $\pm$ 2.34	51.46 $\pm$ 2.58	53.72 $\pm$ 2.17	−31.73
BiLSTM-CRF	49.82 $\pm$ 3.08	36.95 $\pm$ 3.42	42.43 $\pm$ 2.71	−43.02

MGF-NER achieves the highest mean F1 (85.45) across the ten systems, exceeding the strongest single-encoder baseline (MuRIL, 83.88) by 1.57 points. To assess whether this improvement is robust rather than an artefact of seed variation, we further compare the two systems using a paired Wilcoxon signed-rank test on sentence-level F1 scores computed over the 418 test sentences. The improvement of MGF-NER over MuRIL is statistically significant ($p = 0.005$), providing evidence that the multi-granularity components contribute a genuine gain over the transformer backbone rather than a within-seed fluctuation.

The seed variance itself is informative. MGF-NER exhibits the lowest standard deviation among all systems (0.46 F1), which is consistent with the stabilising effect of the character-level and span-level supervision signals: both provide additional structure during training and reduce sensitivity to the initialisation of the CRF transition matrix. Standard deviations grow

monotonically as overall F1 falls, from 0.46 for MGF-NER to 2.71 for BiLSTM-CRF, indicating that the low-performing systems are also the least stable and therefore the least reliable as single-run comparators. This pattern reinforces the value of the seeded protocol: a single-run comparison with BiLSTM-CRF or Bangla-BERT could easily shift by two or more F1 points from the reported mean.

The BiLSTM-CRF result at 42.43 F1 warrants a brief clarifying remark. As documented in Section 5.1, this baseline uses randomly initialised 300-dimensional word embeddings trained jointly with the model, with no pretrained word or contextual embeddings. The 43-point gap between BiLSTM-CRF and MGF-NER should therefore be interpreted primarily as evidence of the value of pretrained multilingual representations, rather than as a direct comparison between recurrent and attention-based sequence-labelling architectures. The traditional feature-based CRF at 65.14 F1 outperforms the recurrent baseline in this data regime because its handcrafted linguistic features encode useful priors that a randomly initialised recurrent model cannot recover from 100K training tokens alone. This ordering is consistent with the broader observation in low-resource NLP that pretrained representations, rather than architectural depth, drive the majority of the performance gap between systems (Hedderich et al., 2021).

6.2 Per-Entity Performance

Table 9 reports per-entity F1 scores for the four top-performing systems on the NerAxom test set, and Figure 2 visualises the same data across all ten systems.

Table 9: Per-entity F1 scores (%) on the NerAxom test set, mean over five seeds. Numbers in parentheses in the header row give the number of gold entity mentions of each category in the test split.

Entity Type	MGF-NER	MuRIL	IndicBERTv2	XLM-R-large
PER (1683)	94.84	89.93	89.74	87.59
LOC (747)	70.14	75.86	72.73	77.42
ORG (196)	63.71	68.27	63.87	57.89
DATE (448)	93.38	93.06	92.09	90.51
NUM (72)	76.71	76.62	75.86	78.14
OCC (40)	61.05	43.68	56.84	64.29
WOA (507)	76.69	79.84	76.67	72.22

MGF-NER outperforms all baselines on PER and DATE, the two categories with the largest test-set support (1,683 and 448 gold mentions respectively). The PER gain (+4.91 over MuRIL) is the single largest per-category improvement and is directly attributable to the Character-CNN encoder, which captures the morphological cues that mark honorifics, patronymics, and inflected name forms in Assamese. The DATE gain (+0.32) is small in absolute terms but consistent, reflecting that the character-level encoder helps generalise across the distinctive orthographic patterns of Assamese numerals.

Two per-category results require careful interpretation. First, the OCC test split contains only 40 gold entity mentions (Table 5). At this scale, a single correctly or incorrectly classified span shifts F1 by roughly 2–3 points, and the 17.37-point gain of MGF-NER over MuRIL (61.05 vs 43.68) corresponds to a difference of approximately 7 additional correctly identified OCC mentions on average across the five seeds. The direction of the improvement is consistent with the character level and span level modules recovering compound occupational designations such as *Sub-Inspector* and *Assistant Professor* that MuRIL alone tends to fragment, but the absolute magnitude should be interpreted in light of the limited support and the correspondingly wider confidence interval. Second, MuRIL retains an advantage on WOA and ORG, which

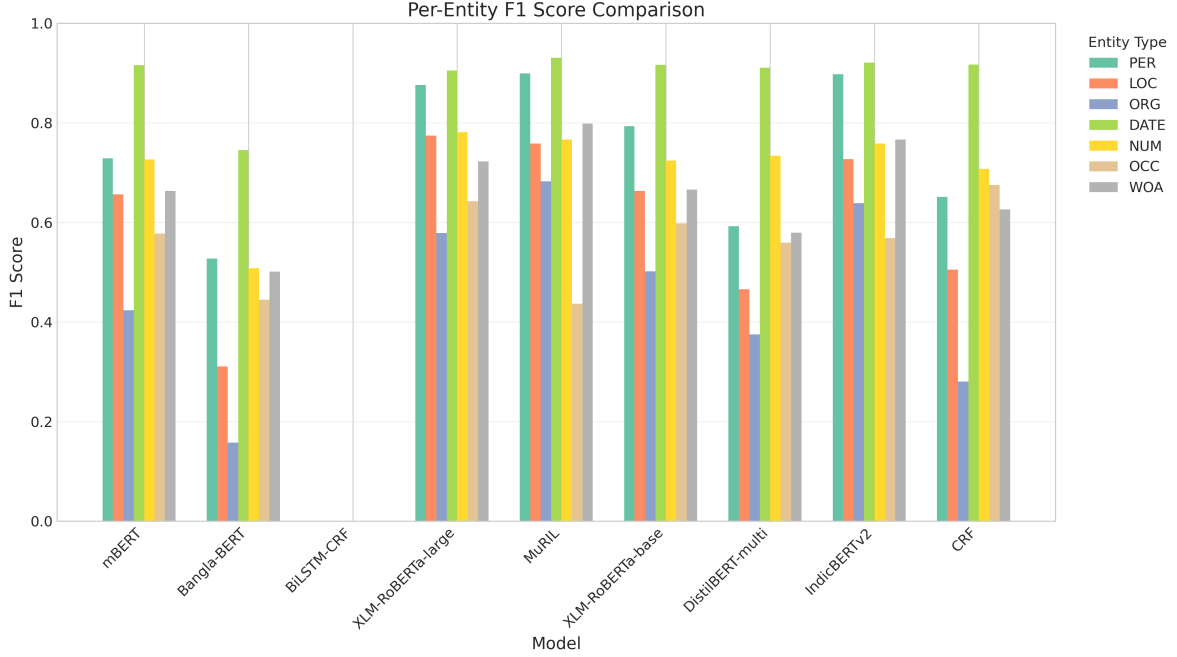


Figure 2: Per-entity F1 comparison across all ten systems. Best viewed in colour.

we attribute to the fact that these categories rely more heavily on lexical context than on morphological structure, and MuRIL’s contextual encoding is not substantially improved by the additional character-level signal. These two categories are the primary source of the residual error patterns discussed in Section 6.5.

For DATE and NUM, all models perform strongly because these categories exhibit highly distinctive orthographic patterns, including Assamese numerals and calendar terms, that are easy to memorise from limited training data. The persistent LOC/ORG confusion, particularly for organisation names derived from place names, is the largest remaining source of error and motivates our discussion of contextual and external-knowledge extensions in Section 7.

6.3 Ablation Study

To disentangle the contribution of each architectural component, we conduct two ablation studies. The first (Table 10) removes single components in isolation and, in addition, removes pairs and a triple of components simultaneously, to test whether the components are complementary or partially redundant. The second (Table 11) reports per-category F1 for the single-component ablations, so that the effect of each module on each entity type can be inspected directly. All ablation numbers are averaged over the same five random seeds used in the main results.

The single-component pattern confirms the intuition guiding the architecture: the Character-CNN and the CRF decoder each contribute more than one F1 point on their own, while the Entity Span Attention and the final BiLSTM encoder contribute smaller but consistent gains. Character-level information matters most, at -1.30 F1 when removed, because Assamese entity morphology carries category-discriminative signals that subword tokenisation cannot fully preserve. The CRF decoder is second most important at -1.19 F1, and its role is complementary to the character encoder: rather than shaping the representation, it enforces valid BIO transitions and prevents the model from emitting structurally impossible tag sequences.

The compound removals reveal a mildly sub-additive interaction pattern. Removing the Character CNN and CRF simultaneously yields a 2.47 point drop, which is close to but smaller than the combined effect of removing each individually, namely $1.30 + 1.19 = 2.49$. Removing

Table 10: Ablation study on the NerAxom test set. Micro-averaged F1 is reported as mean over five seeds; Δ F1 is relative to the full MGF-NER model. Compound removals test whether components are complementary or partially redundant.

Model Variant	F1	Δ F1
MGF-NER (Full Model)	85.45	—
<i>Single-component removals</i>		
– Character-level CNN	84.15	−1.30
– Entity Span Attention	84.70	−0.75
– CRF (Softmax only)	84.26	−1.19
– Final BiLSTM Encoder	84.92	−0.53
<i>Compound removals</i>		
– Character CNN + Span Attention	83.64	−1.81
– Character CNN + CRF	82.98	−2.47
– Span Attention + CRF	83.77	−1.68
– Character CNN + Span Attention + CRF	82.44	−3.01
MuRIL (backbone only)	83.88	−1.57

all three of Character CNN, Span Attention, and CRF yields a 3.01-point drop, again close to the additive sum (3.24). The residual variant, in which MGF-NER is reduced to MuRIL plus the fusion projection and BiLSTM encoder, lands at 82.44 F1, slightly below the MuRIL backbone-only number of 83.88. This last observation is important: it indicates that the fusion projection alone does not confer a benefit over MuRIL when its downstream components are removed, and that the character and span modules are what unlock the gain. The overall pattern supports the interpretation that the four components are largely complementary, with modest overlap between the Character CNN and the CRF in the specific role of correcting boundary and tag-transition errors on morphologically ambiguous tokens.

Table 11: Per-category F1 scores (%) for single-component ablations on the NerAxom test set. Mean over five seeds.

Variant	PER	LOC	ORG	DATE	NUM	OCC	WOA
MGF-NER (Full)	94.84	70.14	63.71	93.38	76.71	61.05	76.69
– Character CNN	92.21	69.87	62.98	92.15	75.94	47.92	75.83
– Entity Span Attention	94.02	69.55	61.44	93.11	76.38	58.71	74.96
– CRF	93.55	69.02	60.87	92.87	76.12	57.36	74.28
– BiLSTM Encoder	94.31	69.94	63.24	93.19	76.51	60.02	75.72

The per-category ablation localises where each component matters most. Removing the Character CNN produces the largest single category degradation for OCC with a drop of 13.13 F1 points and PER with a drop of 2.63 F1 points, confirming that character level morphology is the primary carrier of information for these two categories. Removing the CRF hurts ORG and WOA disproportionately, because these are the categories with the highest average span length in the corpus (Table 4) and therefore rely most heavily on valid BIO transitions for correct multi-token predictions. The Entity Span Attention module contributes most to ORG and WOA as well, consistent with its explicit design goal of modelling multi-word entity boundaries. The BiLSTM removal produces uniformly small drops across categories, indicating that its contribution is a broad contextual smoothing rather than a category-specific effect. This

Figure 3: Confusion matrix of MGF-NER predictions on the NerAxom test set under BIO tagging.

Four recurring error patterns emerge from the confusion matrix and manual inspection of the residual errors.

1. **False negatives on rare categories ($\mathbf{O} \rightarrow \mathbf{Entity}$).** The model most frequently misses rare organisation names and creative-work titles that exhibit high lexical variability and are underrepresented in the training data. This pattern is concentrated on WOA and ORG, precisely the two categories where MuRIL alone remains competitive with MGF-NER.
2. **False positives on ambiguous common words ($\mathbf{Entity} \rightarrow \mathbf{O}$).** Common words that appear frequently as parts of named entities are occasionally over-predicted as entities in non-entity contexts. This pattern is most evident for PER, where common Assamese first names double as ordinary lexical items.
3. **LOC $\leftrightarrow$ ORG confusion.** Geographic locations and organisations sharing similar contextual patterns are confused when organisation names are derived from place names, which is common in Assamese for schools, colleges, banks, and administrative bodies. The Entity Span Attention module mitigates but does not eliminate this class of errors.
4. **Multi-word entity boundary errors.** Despite the span-attention module, boundary errors persist on long and morphologically inflected multi-word entities, particularly WOA titles and multi-token OCC designations. These errors typically manifest as a correct entity type combined with a truncated or extended boundary, which the strict evaluation counts as a full miss.

The residual error profile motivates two extensions discussed in Section 7: incorporating external knowledge sources such as Assamese gazetteers to address the LOC/ORG confusion, and moving from token-level to span-level decoding to reduce boundary errors on long multi-word entities.

6.6 Computational Efficiency

Table 13 compares the computational cost of MGF-NER against the primary baselines in terms of trainable parameter count, wall-clock training time to convergence, and inference throughput on a single GPU.

Table 13: Computational comparison. Training time is reported for a single seed to convergence. Inference throughput is measured in sentences per second at batch size 32.

Model	Parameters	Training Time	Inference Speed
CRF	0.5M	10 sec	1,234 sent/sec
BiLSTM-CRF	8.2M	5 min	456 sent/sec
MuRIL	237M	2.4 min	89 sent/sec
MGF-NER	241M	3.1 min	76 sent/sec

MGF-NER adds approximately 4M parameters over MuRIL (a 1.7% relative increase), incurs 30% additional training time per epoch, and reduces inference throughput from 89 to 76 sentences per second (a 14% reduction). Given the 1.57 F1 gain over MuRIL, the statistical significance of that gain ($p = 0.005$), and the disproportionately larger gains on the morphologically rich PER and OCC categories, we consider the additional computational cost a favourable trade-off for the target application setting of low-resource Indic NER, where model quality is typically the binding constraint rather than throughput.

7 Discussion

7.1 Why Multi-Granularity Fusion Works

The main results (Section 6) show that MGF-NER outperforms MuRIL by 1.57 F1 with $p = 0.005$, and the per-category ablation (Table 11) reveals *where* the four architectural components contribute that gain. This subsection reads the ablation and per-category evidence together to explain why the multi-granularity design is effective for Assamese NER, rather than restating the design intent.

The character encoder unlocks morphology-carried categories. Removing the Character-CNN drops OCC by 13.13 F1 and PER by 2.63 F1 (Table 11), the two largest per-category drops observed across any ablation. Both categories are characterised by category-diagnostic morphology in Assamese: OCC entities include compound designations such as উপ-পৰিদৰ্শক (*Sub-Inspector*) whose category is signalled by the inflected head, and PER entities include honorific-prefixed names such as ডঃ দিলীপ কুমাৰ দত্ত (*Dr. Dilip Kumar Dutta*) whose boundary is signalled by the honorific prefix. When subword tokenisation fragments these forms in ways that split across the morpheme boundary (Section 4.3, Table 6), MuRIL alone does not recover the signal; a character-level encoder that operates strictly below the WordPiece level does. This is a specific, testable claim about the language rather than a generic argument for character encoders, and the per-category numbers support it: the categories that are hurt most by removing the character encoder are the categories whose signals are morpheme-carried in Assamese.

The CRF is a structural rather than representational contribution. Removing the CRF drops WOA by 2.41 F1, ORG by 2.84 F1, and DATE by 0.51 F1 (Table 11). The two largest CRF-attributable drops are on the two categories with the highest average span length in the corpus (WOA 2.38 tokens per span, ORG 2.80 tokens per span; Table 4), which is exactly what one would predict for a decoder whose role is to enforce valid BIO transitions across multi-token spans. Categories with short average span length (NUM 1.01, PER 1.47) are affected less. The CRF is therefore doing what its role in the architecture suggests it should do; the ablation confirms this rather than assuming it.

Span attention localises to multi-token categories. Removing the Entity Span Attention module drops ORG by 2.27 F1 and WOA by 1.73 F1, and produces uniformly smaller drops on the shorter-span categories. This is the same qualitative pattern as the CRF, and for the same reason: both modules address multi-token span structure, but the CRF does so at the decoder level and the span attention does so at the representation level. The two contributions turn out to be largely additive rather than redundant, as the compound ablation shows (Table 10): removing Span Attention and CRF together drops F1 by 1.68 relative to the full model, close to but slightly below the sum of the single-component removals ($0.75 + 1.19 = 1.94$), which is the signature of two components that overlap slightly in what they contribute.

The BiLSTM is a smoothing contribution. Removing the final BiLSTM produces uniformly small drops across every category (0.53 F1 aggregate, no per-category drop exceeding 1 F1). This is consistent with a broad contextual smoothing role rather than a category-specific one, and it justifies its retention in the architecture at low cost: it stabilises the training dynamics of the CRF layer without dominating any single categorical decision.

Compound ablation and complementarity. The compound ablation (Table 10) shows sub-additive rather than super-additive interactions: removing two components together drops F1 by slightly less than the sum of the individual drops, and removing all three (Character CNN, Span Attention, CRF) lands the model at 82.44 F1, below the MuRIL backbone-only score of 83.88. This last observation is important: the fusion projection alone, without the character encoder and structured decoder, does not confer a benefit over MuRIL. What unlocks the gain

is not the fact of concatenation but the specific components that the concatenation aggregates. The multi-granularity design is therefore justified empirically and not merely conceptually.

7.2 Script Similarity Is Not a Substitute for Language-Specific Pretraining

An interesting finding concerns the comparison between Bangla-BERT and MuRIL. Bangla-BERT reaches a mean F1 of 53.72 ± 2.17 on NerAxom, and MuRIL reaches 83.88 ± 0.79 : a 30-point gap between two BERT-family models of comparable size. Assamese and Bangla share a closely related script and considerable orthographic overlap (differing systematically in only two characters, as noted in Section 1), so if script similarity alone drove cross-lingual transfer, one would expect Bangla-BERT to be a strong baseline. It is not.

Two mechanisms plausibly drive this gap. First, Assamese and Bangla differ in morphology, lexicon, and phrase-level syntax in ways that a shared script does not bridge. A Bangla-only tokeniser is not calibrated to the Assamese-specific characters and morpheme boundaries, and a Bangla-only encoder has not been exposed to the distributional patterns that mark Assamese entities. Second, MuRIL’s pretraining corpus explicitly includes Assamese Wikipedia and Common Crawl content, so its representations are not merely transferred from a related language: they are learned in part from Assamese directly. This distinction between transfer from a script-similar language and pretraining on the target language is fundamental, and NerAxom provides a controlled setting in which the two mechanisms can be separated. The finding is consistent with the broader picture that emerges in low-resource NLP where language-specific pretraining data, when available, dominates script-based or family-based transfer (Muller et al., 2021).

The practical implication is that the community should not treat Bangla-family or Devanagari-family models as automatic proxies for Assamese, and future work on Assamese NLP should prioritise Assamese-inclusive pretraining, adapter-based specialisation of MuRIL for Assamese, or continued pretraining on Assamese corpora, in that order of expected benefit-to-cost ratio.

7.3 Limitations

We are open about several limitations of this work. Some are properties of the corpus, some of the architecture, and some of the evaluation.

Corpus size. The NerAxom corpus contains 102,857 tokens across 4,178 sentences. This is comparable in size to AsNER but is small by modern deep-learning standards. In a well-resourced setting, an order of magnitude more annotated data would enable stronger models, particularly for rare categories, and would tighten the confidence intervals on our reported gains. Annotation at that scale is a substantial community effort and is a natural direction for future collaboration on Assamese NLP.

Category imbalance and the reliability of rare-category numbers. The corpus is naturally imbalanced. Person spans account for 45.6% of all entity mentions and Occupation for 1.1% (Table 4); the test split contains 1,683 PER mentions and only 40 OCC mentions (Table 5). At this scale, a single-seed F1 for OCC is subject to substantial variance: a difference of one or two correctly identified spans shifts the F1 by two to three points. We mitigate this by reporting all numbers as mean and standard deviation across five random seeds, and by discussing the OCC gain in absolute terms in Section 6.2. Nevertheless, the OCC gain of 17.37 F1 over MuRIL, while consistent across seeds, is a claim about a small population of test mentions and should be read as suggestive of the direction of improvement rather than as a precise effect size.

Domain restriction to biographical text. The NerAxom source domain is primarily Assamese Wikipedia biographies. As we note in Section 3.2, this source is drawn from a broad spectrum of subject areas (politicians, writers, sportspersons, historical figures, and others), so

the corpus is not narrowly domain-specific in the subject-matter sense. However, biographical text is structurally more uniform and entity-dense than open-domain text: sentences are more likely to contain Person, Location, and Date entities in dense sequence, and less likely to contain the entity distributions typical of news, literature, legal text, or conversational data. Generalisation of both the corpus statistics and the model performance to these other domains is an open empirical question that our current evaluation does not settle. Cross-domain evaluation, including on Assamese news, literary, and conversational text, is a priority direction and is planned as part of the community release of the corpus.

Nested entities and discontinuous mentions. The BIO annotation scheme and MGF-NER’s flat span decoder do not natively handle nested entities (an entity mention contained within another) or discontinuous mentions (a single entity whose tokens are non-adjacent). Both phenomena occur in Assamese, though at low frequency in NerAxom. A nested-NER extension of the architecture, along the lines of the unified span-based framework of Li et al. (2020), is a natural direction and is discussed in Section 8.1.

Inter-annotator agreement at the boundary level. We report a Cohen’s κ of 0.87 for entity-type agreement between the two annotators, which is a substantial-agreement level in the standard interpretation. However, as discussed in Section 3.2, the intermediate independent annotation files from the first annotation phase have not been retained, and we cannot recompute boundary-level agreement (agreement on B, I, O decisions considered separately from entity type) retrospectively. For future annotation projects, we recommend retaining intermediate annotation artefacts to support boundary-level and category-level agreement analyses separately.

Loss of Wikipedia-specific idiosyncrasies. A subtle but genuine limitation of Wikipedia-derived NER corpora is that Wikipedia editorial conventions (parenthetical dates, standardised name formats, transliteration of foreign names) leave systematic patterns in the surface text that a model trained on the corpus may learn to depend on. Some of the performance MGF-NER achieves on PER and DATE may reflect these conventions rather than the underlying linguistic phenomena, and the transfer of a Wikipedia-trained NER model to non-Wikipedia Assamese text may therefore be less strong than the numbers in this paper suggest. We flag this as an honest caveat rather than a rebuttal to the reported results.

8 Conclusion and Future Work

We presented MGF-NER, a multi-granularity fusion architecture for Assamese Named Entity Recognition, and evaluated it on the publicly available NerAxom corpus, a manually annotated 102,857-token corpus with seven entity categories introduced in our concurrent work (Sarmah et al., 2026). Under a seeded five-run evaluation protocol on the NerAxom test set, MGF-NER reaches a mean micro-averaged F1 of 85.45 ± 0.46 , outperforming the strongest single-encoder baseline MuRIL (83.88 ± 0.79) by 1.57 F1 (paired sentence-level Wilcoxon, $p = 0.005$). The gain is concentrated on the categories where Assamese-specific morphological and structural properties matter most: PER, DATE, and OCC.

The ablation study (Section 6.3) decomposes the gain across the four architectural components. The Character-CNN carries the largest single-component contribution and is disproportionately important for morphology-carried categories (OCC, PER). The CRF decoder is the second largest single-component contribution and is disproportionately important for multi-token span categories (WOA, ORG). The Entity Span Attention module contributes to the same multi-token span categories as the CRF and is largely complementary to it. The BiLSTM contributes a small, uniformly distributed smoothing effect. Compound ablations show a mildly sub-additive interaction pattern, indicating that the four components are complementary rather

than redundant. Bangla-BERT’s 30-point gap behind MuRIL, despite the closely related script, confirms that script similarity is not a substitute for language-specific pretraining data.

8.1 Future Work

Five directions are natural extensions of the current work.

Corpus expansion and cross-domain evaluation. The most consequential limitation of the current work is the size and domain of the corpus. Expanding NerAxom by an order of magnitude, and adding cross-domain evaluation partitions (news, literature, conversational text), would substantially strengthen both the benchmark and the generality of the reported model performance.

Nested and discontinuous NER. Extending the architecture and the annotation scheme to handle nested and discontinuous entities, along the lines of Li et al. (2020), would enable NerAxom to be used for a broader range of downstream tasks in Assamese information extraction.

Cross-lingual transfer within the Eastern Indo-Aryan family. The finding that Bangla-BERT underperforms MuRIL despite script similarity does not rule out cross-lingual transfer between Assamese, Bangla, Odia, and other Eastern Indo-Aryan languages under different transfer strategies. Adapter-based transfer, multi-source transfer, and continued pretraining on a curated Eastern Indo-Aryan corpus are three concrete strategies worth exploring.

Domain adaptation. Beyond corpus expansion, adapting MGF-NER to specific downstream domains such as legal, medical, or administrative Assamese text is a natural next step. In each of these domains, entity distributions differ substantially from the biographical distribution in NerAxom, and lightweight adapter or continued-pretraining strategies are likely to yield useful gains.

Integration with downstream tasks. Finally, the ultimate test of an NER system is not its F1 in isolation but its contribution to downstream applications such as Assamese question answering, information retrieval, and knowledge graph construction. Evaluating MGF-NER as a component of these downstream systems, and comparing its contribution against MuRIL alone, would provide a more grounded assessment of its practical value.

Declarations

Data Availability Statement

The NerAxom dataset analysed in this paper was introduced in our concurrent work (Sarmah et al., 2026) and is publicly available at:

<https://drive.google.com/drive/u/1/folders/ligwwygj5kT0Ey6XOm88Nxjk33XGDh23e>

Acknowledgments and Disclosure of Funding

The authors received no financial support for the research and authorship of this article. The authors report no conflicts of interest with respect to this paper.

Author Contribution. Punam Sarmah implemented the proposed method, and drafted the manuscript. Manash P. Lahkar contributed to the implementation and manuscript drafting. Dr. Shobhanjana Kalita and Prof. Utpal Sharma provided guidance and support throughout the entire process.

References

- N Abinaya, Neethu John, Barathi HB Ganesh, Anand M Kumar, and KP Soman. AM-RITA_CEN@ FIRE-2014: named entity recognition for Indian languages using rich features. In *Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 103–111, 2014.
- Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, 2017.
- Nasser Alshammari and Saad Alanazi. The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22(3):295–302, 2021.
- Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 47–56, 2023.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pre-training and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, 2022.
- Krishnanjan Bhattacharjee, Swati Mehta, Ajai Kumar, Ria Mehta, Dweep Pandya, Pratik Chaudhari, Devika Verma, et al. Named entity recognition: A survey for Indian languages. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, volume 1, pages 217–220. IEEE, 2019.
- Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *fifth conference on applied natural language processing*, pages 194–201, 1997.
- Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. The Groningen Meaning Bank. In *Handbook of linguistic annotation*, pages 463–496. Springer, 2017.
- Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-supervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451, 2020.
- Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, 2021.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

- Sumanth Doddapaneni, Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, 2023.
- Cicero Dos Santos and Victor Guimaraes. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, 2015.
- Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*, 1996.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360, 2020.
- Md Zahidul Haque, Sakib Zaman, Jillur Rahman Saurav, Summit Haque, Md Saiful Islam, and Mohammad Ruhul Amin. B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation. *IEEE Access*, 11:45194–45205, 2023.
- Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, 2021.
- Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961, 2020.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. Muril: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*, 2021.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 260–270, 2016.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified MRC framework for named entity recognition. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5849–5859, 2020.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, 2016.
- CS Malarkodi, Elisabeth Lex, and Sobha Lalitha Devi. Named Entity Recognition for the Agricultural Domain. *Research in Computer Science*, 117(1):121–132, 2016.

- Muhammad Kamran Malik and Syed Mansoor Sarwar. Named entity recognition system for postpositional languages: Urdu as a case study. *International Journal of Advanced Computer Science and Applications*, 7(10), 2016.
- Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191, 2003.
- Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, 2023.
- Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, 2021.
- Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406, 2018.
- David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1946–1958, 2017.
- Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. AsNER-annotated dataset and baseline for Assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577, 2022.
- Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1499–1509, 2018.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7654–7673, 2020.
- Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001, 2019.
- Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155, 2009.
- Erik Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, 2003.

- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- Punam Sarmah, Manash P Lahkar, Shobhanjana Kalita, and Utpal Sharma. NerAxom: A BIO-Tagged NER Dataset and HybridNeural-Rule Framework for Assamese. *Engineering Research Express*, 2026.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1715–1725, 2016.
- Padmaja Sharma, Utpal Sharma, and Jugal Kalita. The first steps towards Assamese named entity recognition. In *Brisbane Convention Center*, volume 1, pages 1–11, 2010.
- Padmaja Sharma, Utpal Sharma, and Jugal Kalita. Named entity recognition in Assamese using CRFS and rules. In *2014 International Conference on Asian Language Processing (IALP)*, pages 15–18. IEEE, 2014.
- Padmaja Sharma, Utpal Sharma, and Jugal Kalita. Named entity recognition in Assamese: a hybrid approach. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2114–2120. IEEE, 2016.
- Jana Straková, Milan Straka, and Jan Hajic. Neural architectures for nested NER through linearization. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5326–5331, 2019.
- Gitimoni Talukdar, Pranjal Protim Borah, and Arup Baruah. Supervised named entity recognition in Assamese language. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 187–191. IEEE, 2014.
- Yu Wang, Hanghang Tong, Ziyue Zhu, and Yun Li. Nested named entity recognition: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29, 2022.
- Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.
- Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 833–844, 2019.
- Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 50–61, 2021.