Journal Title
XX(X):1–14

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Neuro-Symbolic Relation Extraction in Agglutinative Languages: A Morphology-Aware Graph-Based Framework

Laishram Jimmy[1] and Biri Arun[2]

Abstract

Entity relationship extraction (ERE) is a core task in natural language processing, enabling structured knowledge extraction and downstream reasoning. However, it remains particularly challenging in low-resource, morphologically rich languages, where limited data, absence of reliable parsers, and complex affix-driven grammatical structures hinder conventional neural and dependency-based approaches.

In such languages, relational cues are often encoded through suffixes rather than word order, making them difficult to capture using token-level representations. To address this, we propose a morphology-aware graph-based framework that constructs heterogeneous sentence graphs from morphologically segmented text. The model treats grammatical affixes as explicit graph nodes, enabling direct modeling of linguistic roles through structured message passing.

To enhance interpretability and robustness, we integrate a neuro-symbolic classification layer based on the Tsetlin Machine (TM). Instead of acting as a conventional classifier, the TM learns propositional rules over symbolic features derived from graph structure and binarized R-GCN node embeddings. This allows the model to translate distributed neural representations into compact, human-interpretable patterns that capture entity-type interactions, morphological cues, and structural dependencies, leading to more stable and transparent relation prediction.

We evaluate the proposed framework on a newly introduced dataset for extracting Manipuri relation called **MERED**. The model achieves 91.72% accuracy and 91.16% Macro-F1, significantly outperforming strong neural baselines. These results demonstrate that combining morphology-aware graph modeling with symbolic reasoning provides an effective, interpretable, and data-efficient solution for relation extraction in low-resource agglutinative languages.

Keywords

Entity Relationship Extraction, Low-resource NLP, Morphologically Rich Languages, Graph Neural Networks, Tsetlin Machines

Introduction

Entity Relationship Extraction (ERE) is a core task in Natural Language Processing (NLP) that enables the transformation of unstructured text into structured knowledge. Accurate identification of entities and their semantic relations underpins applications such as knowledge graph construction, information retrieval, event understanding, and decision-support systems (Zhao et al. 2023).

While recent advances in deep learning and pretrained language models have significantly improved ERE, these gains are largely confined to high-resource languages such as English and Chinese (Zhao et al. 2023; Li and Ji 2016; Zhang et al. 2018). In contrast, low-resource languages continue to face fundamental challenges due to limited annotated data, lack of linguistic tools, and the absence of reliable syntactic parsers (Philip et al. 2020; Lahoti et al. 2022). These constraints limit the applicability of conventional sequence-based and parser-dependent approaches.

The difficulty is further amplified in agglutinative and morphologically rich languages, where grammatical relations are not expressed through fixed word order but are encoded through affixes and case markers attached to lexical units (Bickel 2015). Consequently, relational information is distributed across subword structures and often misaligned with token boundaries, making word-level representations inadequate (Lahoti et al. 2022; Bolucu et al. 2021). Additionally, flexible word order and pro-drop constructions weaken positional cues, increasing ambiguity in relation extraction (Comrie 1989b; Huang et al. 2018; Bisazza and Federico 2014).

These linguistic characteristics introduce several challenges: (i) relations are encoded at the subword level, requiring models to capture affix-level information (Haspelmath 2002; Tsarfaty et al. 2010b); (ii) free word order reduces the reliability of positional features (Comrie 1989a; Bickel 2015); (iii) implicit arguments increase ambiguity due to omitted entities; and (iv) morphological ambiguity arises when the same affix

[1]Manipur Technical University, Imphal, Manipur, India
[2]National Institute of Technology, Jote, Arunachal Pradesh, India

Corresponding author:
Laishram Jimmy
Email: jimmy_l@mtu.ac.in

expresses multiple semantic roles depending on context (Tsarfaty et al. 2010b; Cotterell et al. 2018b). Together, these factors make dependency parsing unreliable and limit the effectiveness of traditional neural models in low-resource settings (He 2023).

To address these challenges, we propose a morphology-aware graph-based framework that reconceptualizes relation extraction at the morpheme level. Instead of treating morphology as auxiliary features, grammatical affixes are modeled as first-class nodes in a heterogeneous sentence graph. Lexical roots, affixes, and entity mentions are explicitly represented, and their interactions are encoded through typed edges capturing morphological, sequential, and semantic relationships. This design enables grammatical role information to directly influence representation learning through relational graph convolutional networks (R-GCNs) (Kipf and Welling 2017; Zhou et al. 2020; Wu et al. 2021a).

Building on this representation, we introduce a neuro-symbolic reasoning layer using the Tsetlin Machine (TM) (Granmo 2018). Unlike conventional neural classifiers, the TM learns propositional logic clauses over symbolic features derived from graph structure and binarized node embeddings, effectively bridging distributed neural representations with explicit rule-based reasoning. This allows the model to transform learned node embeddings into compact, interpretable patterns that capture entity-type interactions, morphological cues, and structural dependencies. The resulting framework performs pattern-level reasoning rather than purely statistical classification, leading to improved robustness and transparency. Notably, the learned clauses reveal consistent relation-specific structures. For example, `AFFECTS` is frequently associated with event–group interaction patterns which is further analyzed in Section .

We evaluate the proposed approach on **MERED** (Manipuri Entity Relationship Extraction Dataset), a newly constructed dataset for Manipuri, a low-resource agglutinative language. Experimental results demonstrate significant improvements over sequence-based, transformer-based, and graph-based baselines, highlighting the effectiveness of combining morphology-aware graph modeling with interpretable neuro-symbolic reasoning.

Review of Literature

Entity Relationship Extraction (ERE) is a core task in Natural Language Processing (NLP), aimed at identifying semantic relations between entities mentioned in text. Reliable ERE systems support a wide range of applications, including knowledge graph construction, information retrieval, and question answering (Zhao et al. 2023; Bach et al. 2004; Yao et al. 2019).

Recent advances in ERE have been largely driven by deep neural models, particularly transformer-based architectures, which achieve strong performance in high-resource languages such as English and Chinese (Devlin 2019; Wang et al. 2020). However, these models depend heavily on large annotated datasets and well-developed linguistic resources, limiting their applicability in low-resource settings (Philip et al. 2020). This limitation becomes more pronounced in morphologically rich and agglutinative languages, where linguistic structure differs significantly from widely studied languages such as English (Tsarfaty et al. 2010a).

Many Indian languages, including Manipuri, Tamil, Telugu, and Marathi, exhibit agglutinative or morphologically rich characteristics. In these languages, grammatical relations are encoded through affixes, case markers, and postpositions attached to lexical units rather than through fixed word order (Comrie 1989b; Devi and Singh 2019). Consequently, relational information is distributed across subword units and does not align with surface tokens, making it difficult for sequence-based models to accurately capture entity boundaries and semantic roles (Tsarfaty et al. 2010a). Manipuri (Meiteilon), in particular, demonstrates this complexity, where a single word may encode multiple grammatical functions, leading to ambiguity when treated as a single token (Devi and Singh 2019). Similar challenges have been observed across other Indian languages, where standard NLP pipelines fail to generalize due to mismatched linguistic assumptions (Lahoti et al. 2022).

To address these challenges, morphological segmentation has been widely explored as a preprocessing strategy. By decomposing words into roots and affixes, segmentation reduces sparsity and improves representation learning (Creutz and Lagus 2007; Sennrich et al. 2016). Prior studies have shown that segmentation improves performance in tasks such as named entity recognition and part-of-speech tagging in low-resource languages (Singh and Sharma 2021; Garg and Sharma 2022). However, in most existing approaches, segmentation is used only as a preprocessing step, and the extracted morphemes are not directly incorporated into the model architecture. As a result, while segmentation improves lexical coverage, it does not fully exploit the grammatical signals encoded in morphology, which are crucial for relation extraction (Cotterell et al. 2018a).

Graph-based models have emerged as an effective alternative to sequence-based approaches for capturing structural dependencies. Graph Neural Networks (GNNs) represent text as graphs, allowing information to propagate across non-adjacent elements through message passing (Wu et al. 2021b; Zhou et al. 2020). Graph Convolutional Networks (GCNs), in particular, have demonstrated strong performance in relation extraction when combined with syntactic dependency structures (Kipf and Welling 2017; Zhang et al. 2018). These models are well-suited for morphologically rich languages, as they can capture non-linear relationships and long-range dependencies beyond linear token sequences (Battaglia et al. 2018). Recent studies have shown that morphology-aware graph models outperform transformer-based approaches in low-resource settings, particularly when syntactic information is incomplete or noisy (He et al. 2023; Ji et al. 2022).

Despite these advances, most graph-based approaches still rely on dependency parsers to construct the underlying graph structure. This dependency poses a significant limitation in low-resource languages, where annotated treebanks are scarce and parsing performance is often unreliable (Zeman et al. 2017; Kanerva and Ginter 2018). Furthermore, existing methods typically treat morphological information as auxiliary features rather than as core structural components of the graph, limiting their ability to capture fine-grained grammatical relations (Cotterell et al. 2018a).

In parallel, neuro-symbolic approaches have gained attention for improving interpretability and robustness in NLP systems. The Tsetlin Machine (TM) is a rule-based learning framework that models patterns using propositional logic and has demonstrated competitive performance in classification tasks while maintaining high interpretability (Granmo 2018; Abeyrathna et al. 2021). Unlike deep neural models, which rely on dense representations, TM-based approaches learn explicit logical clauses, making them suitable for low-resource settings where data is limited and interpretability is critical. Recent work suggests that combining neural representations with symbolic reasoning can improve both generalization and transparency, particularly in structured prediction tasks (d'Avila Garcez et al. 2019).

However, the integration of morphology-aware graph representations with neuro-symbolic reasoning remains largely unexplored. Most existing studies either focus on neural models without interpretability or symbolic models without leveraging rich linguistic structure. This gap highlights the need for a unified framework that can combine the strengths of both paradigms.

In this work, we address these limitations by proposing a parser-independent, morphology-aware graph framework that explicitly models grammatical roles through morpheme-level representations. By treating role-marking affixes as first-class graph nodes and integrating a neuro-symbolic classification layer based on the Tsetlin Machine, the proposed approach enables both accurate and interpretable relation extraction in low-resource, morphologically rich languages, with a particular focus on Manipuri.

System Design

The proposed framework is motivated by the observation that, in agglutinative languages, important relational cues are often carried by morphemes rather than surface words. To capture this structure, each sentence is converted into a heterogeneous graph containing lexical roots, grammatical affixes, and entity mentions. The graph is encoded using a two-layer relational graph convolutional network (R-GCN), and relation prediction is performed using a Tsetlin Machine (TM) over graph-derived symbolic and binarized neural features. Figure 1 shows the overall pipeline.

Morphology-Aware Heterogeneous Graph

Each sentence is segmented into lexical stems and grammatical suffixes using a greedy suffix-based strategy over a curated inventory of Manipuri morphemes. Instead of treating suffixes as auxiliary tags, the framework promotes them to graph nodes so that grammatical information can directly participate in message passing (Haspelmath 2002; Bickel 2015).

We construct a heterogeneous graph $G = (V, E, \mathcal{R})$, where V contains root nodes, affix nodes, and entity nodes. Root nodes represent lexical stems, affix nodes represent grammatical markers (e.g., genitive, locative), and entity nodes represent annotated spans. The edge schema includes `HAS_AFFIX`, `NEXT_TOKEN`, `SPAN_OF`, and `SELF_LOOP`. Such heterogeneous representations allow different linguistic interactions to be modeled explicitly, improving relational reasoning (Zhu et al. 2019; Schlichtkrull et al. 2018a; Zhou et al. 2020).

Node initialization is type-specific. Root nodes combine lexical, POS, and NER information; affix nodes encode suffix form and grammatical role; and entity nodes combine entity type with pooled span-level context. Lexical embeddings are initialized using FastText to better handle sparsity in morphologically rich languages (Bojanowski et al. 2017), while POS and NER features provide additional syntactic and semantic grounding (Rahimi et al. 2019).

Node representations are learned using a two-layer R-GCN, which performs relation-specific message passing over the heterogeneous graph:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right), \quad (1)$$

where $\mathcal{N}_i^r$ denotes neighbors under relation r, $W_r^{(l)}$ are relation-specific transformations, and σ is a nonlinear activation. R-GCNs are well-suited for heterogeneous graphs as they preserve semantic differences across relation types (Schlichtkrull et al. 2018a; Zhu et al. 2019). Two layers provide a balance between local morphological propagation and broader contextual reasoning while avoiding over-smoothing (Li et al. 2018).

As illustrated in Figure 2, root, affix, and entity nodes exchange information through relation-specific edges, allowing grammatical markers and entity context to jointly influence node representations.

The final node representation is $z_i = h_i^{(2)}$.

Neuro-Symbolic Relation Classification

Given entity mentions, candidate head–tail pairs are generated using lightweight type, distance, and graph connectivity constraints (Miwa and Bansal 2016; Zeng et al. 2018). For each entity e_k, a contextual representation is obtained by mean pooling:

$$g_k = \frac{1}{|S_k|}\sum_{v_i\in S_k} z_i. \quad (2)$$

For a pair (e_h, e_t), the representation is:

$$r_{h,t} = [g_h; g_t]. \quad (3)$$

Instead of using a purely neural classifier, the framework converts graph-derived evidence into binary

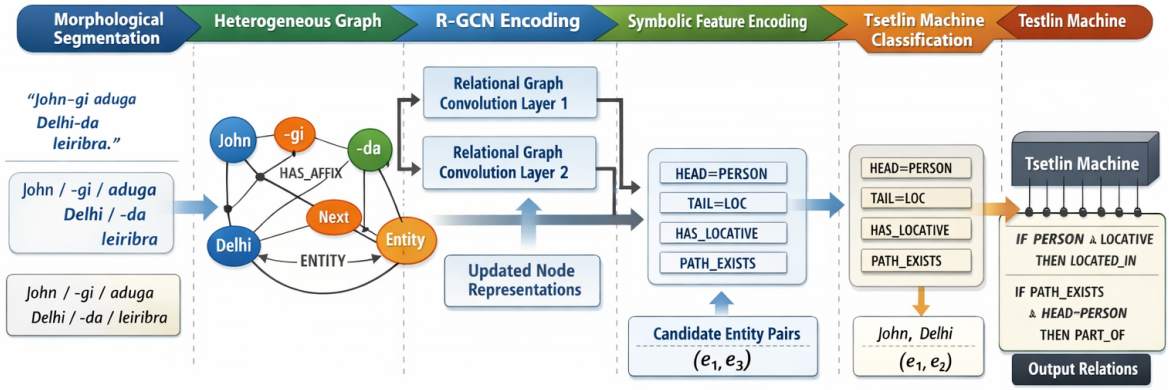

Figure 1. Overall architecture of the proposed neuro-symbolic relation extraction framework. Morphologically segmented text is converted into a heterogeneous graph, encoded using an R-GCN, and classified using a Tsetlin Machine.

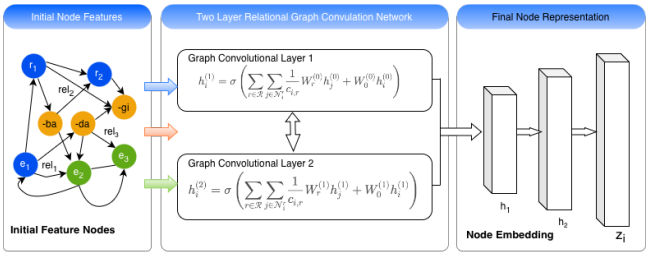

Figure 2. Internal workflow of the R-GCN encoder. Root, affix, and entity nodes interact through relation-specific message passing to produce contextual representations.

symbolic features capturing entity types, structural connectivity, morphological markers, and binarized embedding signals. These features are passed to a Tsetlin Machine (TM), which learns propositional clauses of the form:

$$C_j(\mathbf{x}) = \bigwedge_{k \in I_j} x_k \ \wedge \ \bigwedge_{l \in \bar{I}_j} \neg x_l, \tag{4}$$

where I_j and $\bar{I}_j$ denote the sets of included and excluded features, respectively.

In practice, each clause corresponds to an interpretable conjunction of relational conditions over the input feature space. For example, a learned clause for the relation `AFFECTS` can be expressed as:

$$\texttt{TAIL\_TYPE\_GROUP} \wedge \texttt{PAIR\_EVENT\_GROUP} \wedge \texttt{EMB\_POS\_8} \wedge \neg\texttt{PAIR\_LOC\_LOC} \wedge \neg\texttt{TAIL\_TYPE\_TITLE} \rightarrow \texttt{AFFECTS}. \tag{5}$$

This clause illustrates how the TM combines different feature types: (i) *symbolic features* such as entity-type compatibility (`PAIR_EVENT_GROUP`), (ii) *structural constraints* (implicitly captured through graph-derived features), and (iii) *binarized neural features* (e.g., `EMB_POS_8`), while (iv) excluding incompatible patterns through negated literals (e.g., $\neg$`PAIR_LOC_LOC`).

Importantly, the positive literals define the core relational pattern, whereas the negative literals act as constraints that refine the decision boundary by eliminating competing interpretations. This results in compact, human-interpretable rules that align with linguistic structure.

The TM is trained using supervised Type I and Type II feedback mechanisms (Granmo 2018). Let y denote the ground truth relation and $\hat{y}$ the predicted class. For a clause $C_j(\mathbf{x}) \in \{0, 1\}$, feedback is applied as follows:

- **Type I Feedback (Reinforcement):** applied when $y = \hat{y}$, encouraging inclusion of literals that contribute to correct predictions:

$$P(\text{include } x_k) \uparrow \quad \text{if } C_j(\mathbf{x}) = 1 \wedge y = \hat{y} \tag{6}$$

- **Type II Feedback (Suppression):** applied when $y \neq \hat{y}$, encouraging inclusion of negated literals to prevent false positives:

$$P(\text{include } \neg x_k) \uparrow \quad \text{if } C_j(\mathbf{x}) = 1 \wedge y \neq \hat{y} \tag{7}$$

In the multi-class setting, feedback is applied positively for the target relation and negatively for competing classes, enabling the model to learn discriminative clause sets.

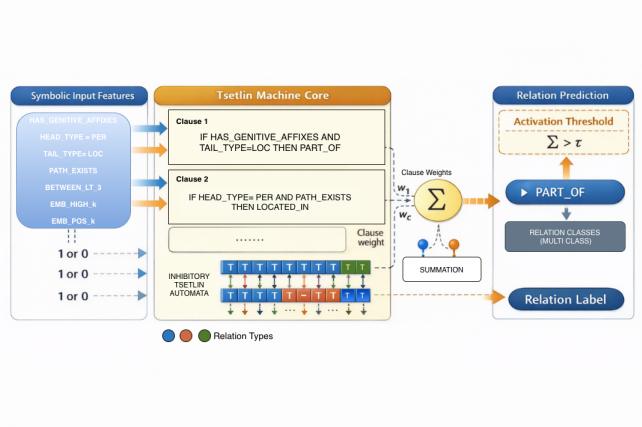

Figure 3. Internal working of the Tsetlin Machine classifier. Binary features derived from graph structure, morphology, and binarized R-GCN embeddings are updated via Type I and Type II feedback to learn propositional clauses, which are combined through voting for final relation prediction.

For example, a learned clause for the relation `AFFECTS` can be expressed as:

$$\texttt{PAIR\_EVENT\_GROUP} \wedge \texttt{TAIL\_TYPE\_GROUP} \wedge \neg\texttt{PAIR\_LOC\_LOC} \wedge \neg\texttt{TAIL\_TYPE\_TITLE} \rightarrow \texttt{AFFECTS}. \quad (8)$$

Here, Type I feedback reinforces the core pattern (`PAIR_EVENT_GROUP`), while Type II feedback introduces exclusion constraints (e.g., $\neg$`PAIR_LOC_LOC`) to suppress incorrect activations. This interplay between reinforcement and suppression allows the TM to learn compact, interpretable rules that align with relational structure.

This feedback-driven clause construction effectively transforms distributed graph embeddings into explicit logical rules, allowing the model to retain the representational power of neural encoders while achieving interpretable and robust decision-making (Abeyrathna et al. 2021; d'Avila Garcez et al. 2019). The internal workflow of clause learning and voting is illustrated in Figure 3.

Dataset

MERED: Dataset Overview

We introduce **MERED** (Manipuri Entity Relationship Extraction Dataset), a manually annotated corpus designed for morphology-aware relation extraction in low-resource agglutinative settings. The dataset explicitly captures morphological, syntactic, and semantic signals, which are critical in languages where grammatical relations are encoded through affixes rather than fixed word order. A sample version is available at: https://www.kaggle.com/datasets/jimmylais/mered-dataset.

Annotation Schema

MERED provides multi-layer annotations at the morpheme level, including morphological segmentation, POS, NER, morpheme roles, entity spans, and relation labels. Words are decomposed into root and affix units using a morphology-aware segmentation strategy, enabling fine-grained modeling of grammatical roles.

Table 1 summarizes the annotation schema with examples.

Relation Inventory

MERED defines a compact set of linguistically grounded relations aligned with morphological cues such as case markers, genitives, and locatives. These relations capture both structural and semantic interactions between entities.

Table 2 illustrates the mapping between linguistic relations and MERED labels.

Utility and Transferability

MERED combines morphological decomposition with multi-layer linguistic annotation, enabling robust modeling of morpho-syntactic and semantic relations that are difficult to capture using word-level representations.

Importantly, the dataset design is language-agnostic. It is based on universal linguistic phenomena such as case marking, affixal role encoding, and morpheme–root attachment rather than dependency structures (Comrie 1989b; Haspelmath 2010).

As a result, the methodology is directly transferable to other low-resource agglutinative languages such as Uyghur, Bodo, Santali, and Kokborok (He 2023; Basumatary 2014; Murmu 2015; Debbarma 2017). Unlike parser-dependent approaches, our framework relies only on morphological segmentation and local attachment rules, making it practical in extremely low-resource scenarios (Nivre 2016).

Thus, MERED serves both as a benchmark dataset for Manipuri and as a general blueprint for morphology-aware relation extraction in under-represented languages.

Evaluation Methodology

Relation extraction is evaluated by exact matching of head entity, tail entity, and relation label.

Metrics

To address class imbalance, we report:

- **Micro-F1**: overall performance,
- **Macro-F1**: performance across classes,
- **Precision/Recall**: error analysis.

Pipeline-Level Evaluation

Given the hybrid architecture, evaluation is performed at multiple levels:

- **Clause Interpretability**: analysis of TM clauses to assess linguistic alignment.
- **End-to-End Performance**: full R-GCN + TM pipeline.
- **Graph Encoding**: R-GCN embeddings with softmax classifier.
- **Symbolic Reasoning**: TM over graph-derived features.

Table 1. Annotation schema and examples in the MERED dataset

Field	Description	Example (English Transliterated)	Annotation Output
Sentence	Original Manipuri sentence	*Tombana cycle bazaardagi leiye*	N/A
Morphological Segmentation	Root–affix decomposition (Singh and Sharma 2021; Garg and Sharma 2022)	*Tomba* + *-na*, *cycle*, *bazaar* + *-da* + *-gi*, *lei* + *-ye*	[ROOT, AFFIX, ROOT, ROOT, AFFIX, AFFIX, ROOT, AFFIX]
NER Tags	BIO-style entity labels (Ma et al. 2022)	*Tomba* → B-PER; *bazaar* → B-LOC	[B-PER, O, O, B-LOC, O, O, O, O]
POS Tags	Part-of-speech categories (Devi and Singh 2019)	NOUN, CASE, NOUN, NOUN, CASE, CASE, VERB, TENSE	[NOUN, CASE, NOUN, NOUN, CASE, CASE, VERB, TENSE]
Morpheme Roles	Root vs. grammatical affix (He et al. 2023)	*-na* → subject marker; *-da* → locative; *-gi* → genitive	[ROOT, AFFIX, ROOT, ROOT, AFFIX, AFFIX, ROOT, AFFIX]
Entity Pairs	Entity mentions in relation (Ji et al. 2022)	(*Tomba*, *bazaar*)	(e_1, e_2)
Relation Type	Semantic relation label (Wu et al. 2021a)	located-in	located-in

Table 2. Mapping between linguistic relations in Manipuri and semantic relation labels used in MERED

Linguistic Relation	Description	Example (Transliterated)	Morphological Cue	Mapped MERED Label
AGENT-ACTION	Entity performing an action	Jaykumar+*na* speech give	Agent marker *-na*	`AFFECTS`, `USES`
PATIENT-OF	Entity affected by an action	pauche+*bu* give	Object marker *-bu*	`AFFECTS`, `TREATS`
LOCATION-OF	Event or entity occurring at a location	Imphal+*da* event	Locative marker *-da*	`LOCATED_IN`
SOURCE-OF	Origin or source of an event/action	wan+*dagi* come	Ablative marker *-dagi*	`ASSOCIATED_WITH`
POSSESSION	Ownership or part-whole relation	council+*gi* member	Genitive marker *-gi*	`PART_OF`
AFFILIATION	Membership or institutional relation	committee+*gi* secretary	Genitive marker *-gi*	`AFFILIATED_WITH`
APPOINTMENT	Assignment to a role or position	secretary *oiba* person	Copula/role marker *oiba*	`HOLDS_TITLE`
QUANTITY	Numerical or value-based relation	lupa 50	Numeric association (no suffix)	`HAS_VALUE`
TEMPORAL	Time-related relation	tha *amda* event	Temporal marker *amda*	`ASSOCIATED_WITH`
COORDINATION	Conjunction or co-occurrence of entities/events	amadi link	Conjunctive marker *amadi*	`ASSOCIATED_WITH`, `PARTICIPATES_IN`

All evaluations are conducted on filtered candidate pairs to ensure balanced fine-grained classification.

Baselines

We compare against representative models:

- BiLSTM (Huang et al. 2015)
- CNN–BiLSTM (Zeng 2014)
- XLM-R (Conneau 2020)
- Dependency-based GCN (Zhang et al. 2018)
- R-GCN (neural baseline)

These baselines cover sequential, transformer-based, and graph-based approaches, highlighting the impact of morphology-aware modeling.

Hyperparameter Configuration

The framework consists of a graph encoder and a Tsetlin Machine (TM) classifier.

Graph Encoder

Table 3 summarizes R-GCN settings.

We use a two-layer R-GCN to balance local morphological propagation and global context while

Table 3. Hyperparameters for the heterogeneous R-GCN encoder

Hyperparameter	Value
Number of R-GCN Layers	2
Root Embedding Dimension	100
POS Embedding Dimension	16
NER Embedding Dimension	16
Affix-role Embedding Dimension	16
Node-type Embedding Dimension	8
Hidden Dimension	128
Dropout Rate	0.2
Activation Function	ReLU
Neighborhood Aggregation	Mean
Optimizer	Adam
Learning Rate	1×10^{-3}
Loss Function	Weighted Cross-Entropy
Negative Sampling Ratio	3:1
Training Epochs	10
Batching Strategy	Sentence-wise graphs

avoiding over-smoothing (Li et al. 2018). Root embeddings are initialized using FastText (Bojanowski et al. 2017), while POS, NER, and affix-role embeddings provide complementary signals (Rahimi et al. 2019).

Weighted cross-entropy and negative sampling (3:1) address class imbalance (Zeng et al. 2015). Training uses Adam optimization (Kingma and Ba 2014) with dropout regularization (Srivastava et al. 2014).

Tsetlin Machine

Table 4 summarizes TM settings.

Table 4. Hyperparameters for the Tsetlin Machine classifier

Hyperparameter	Value
TM Variant	Multi-class Tsetlin Machine
Number of Clauses	2000-5000
Threshold (T)	500-800
Specificity Parameter (s)	5.0
Clause Polarity	Positive and Negative
Input Type	Binary symbolic features
Feature Source	Graph-derived pair features
Training Epochs	100
Voting Scheme	Clause summation
Output Space	Reduced relation label set

The TM operates on binary graph-derived features and learns propositional logic clauses via reinforcement feedback (Granmo 2018). Clause count controls capacity, while T and s regulate decision thresholds and specificity (Abeyrathna et al. 2021).

Pipeline Configuration

Table 5 summarizes pipeline settings.

Table 5. Pipeline-level configuration

Configuration	Value
Morphological Segmentation Strategy	Constrained subword segmentation
Graph Node Types	Root, Affix, Entity
Graph Edge Types	`HAS_AFFIX`, `NEXT_TOKEN`, `SPAN_OF`, `SELF_LOOP`
Entity Representation	Mean pooling
Candidate Pair Generation	Ordered entity pairs
Candidate Pruning	Type, distance, graph connectivity
Relation Label Space	Reduced semantic label inventory

The pipeline combines morphology-aware graph construction, R-GCN encoding, and symbolic classification, enabling expressive yet interpretable relation extraction (d'Avila Garcez et al. 2019).

Results and Discussion

Overview

The proposed framework is evaluated under two complementary settings to comprehensively assess its effectiveness for relation extraction in low-resource, morphologically rich languages.

We first analyze the learned decision patterns of the Tsetlin Machine (TM) to understand how the model captures relational structure. By extracting and inspecting logical clauses, we examine the role of structural features, morphology-aware cues, and neural embedding signals in relation prediction. This analysis provides insight into the model's reasoning process and validates whether linguistically meaningful patterns are learned.

Next, we evaluate the model on the full end-to-end relation extraction task, where it must identify and classify relations among all candidate entity pairs, including those with no relation. This setting reflects real-world deployment conditions but is inherently challenging due to severe class imbalance.

Finally, we evaluate the model on fine-grained relation classification, where the model is trained and tested only on entity pairs that express valid relations. This setting isolates the model's ability to distinguish between different semantic relations without being affected by the dominant negative class.

Interpretability of Learned Tsetlin Clauses

A key advantage of the proposed neuro-symbolic framework is that the final classification stage remains interpretable. The features used by TM is explained in Table 6. Unlike purely neural classifiers, the Tsetlin Machine (TM) learns explicit logical clauses over binary features derived from entity-pair types, graph structure, morphological cues, and binarized R-GCN embeddings. This makes it possible to inspect the learned decision patterns and relate them directly to linguistic behavior in Manipuri.

Clause Structure. The learned TM clauses consist of conjunctions of binary literals. These literals include:

- entity-type and pair-type indicators such as `HEAD_TYPE_PER`, `TAIL_TYPE_GROUP`, and `PAIR_EVENT_GROUP`,
- graph and positional features such as `SAME_SENTENCE`, `GRAPH_CONNECTED`, `HEAD_BEFORE_TAIL`, and distance based indicators,
- morphology-aware features such as `HAS_GENITIVE`, `HAS_LOCATIVE`, `HEAD_HAS_*`, and `BETWEEN_HAS_*`,
- binarized embedding features such as `EMB_HIGH_k` and `EMB_POS_k`, which act as neural refinements over the symbolic feature space.

Although some clauses contain many embedding-based literals, the most informative part of the learned logic is often the interpretable symbolic core. In practice, the embedding literals help refine decision boundaries, while the pair-type, entity-type, and morphology-aware literals provide the human-readable explanation.

Learned Clause Patterns The learned clauses reveal that relation prediction is governed by a small set of recurring structural, semantic, and morphology-aware patterns. Rather than learning arbitrary rules, the Tsetlin Machine consistently captures relation-specific interactions between entity types, refined by graph connectivity and contextual constraints.

Structure-driven relations. Relations such as `AFFECTS`, `LOCATED_IN`, and `PARTICIPATES_IN` are primarily governed by entity-type interactions and graph structure:

$$\texttt{PAIR\_EVENT\_GROUP} \wedge \texttt{GRAPH\_CONNECTED} \rightarrow \texttt{AFFECTS},$$

$$\texttt{PAIR\_EVENT\_LOC} \vee \texttt{TAIL\_TYPE\_LOC} \rightarrow \texttt{LOCATED\_IN},$$

$$\texttt{PAIR\_GROUP\_EVENT} \wedge \texttt{TAIL\_TYPE\_EVENT} \rightarrow \texttt{PARTICIPATES\_IN}.$$

These relations rely on structural connectivity rather than explicit morphological cues.

Type-driven relations. Relations such as `TREATS` and `USES` are dominated by entity-type compatibility:

$$\texttt{HEAD\_TYPE\_CHEMICAL} \wedge \texttt{TAIL\_TYPE\_SEED} \rightarrow \texttt{TREATS},$$

$$\texttt{PAIR\_GROUP\_VEHICLE} \vee \texttt{TAIL\_TYPE\_CONTROL\_MEASURE} \rightarrow \texttt{USES}.$$

These patterns indicate that domain-specific semantics can be captured through compact symbolic rules.

Domain-specific relations. The relation `MIXED_WITH` exhibits highly stable patterns:

$$\texttt{PAIR\_CHEMICAL\_RESOURCE} \rightarrow \texttt{MIXED\_WITH}.$$

Such relations are learned with minimal ambiguity due to strong semantic constraints.

Multi-pattern relations. The relation `OPPOSES` is modeled through multiple sub-patterns:

$$\texttt{PAIR\_PER\_PER} \vee \texttt{PAIR\_ORG\_LOC} \rightarrow \texttt{OPPOSES}.$$

This shows that the model captures heterogeneous interaction types rather than a single fixed rule.

Fallback relation. In contrast, `ASSOCIATED_WITH` lacks a strong positive signature and is defined through weak contextual and exclusion-based patterns:

$$\neg\texttt{PAIR\_TITLE\_TITLE}, \neg\texttt{PAIR\_GROUP\_GROUP}, \neg\texttt{HEAD\_TYPE\_DATE}.$$

This indicates that it functions as a *fallback relation*, capturing general co-occurrence when no stronger structural or semantic pattern is present.

Summary. Overall, the learned clauses demonstrate that relation extraction is driven by a combination of: (i) entity-type compatibility, (ii) graph connectivity, and (iii) selective morphological cues. The Tsetlin Machine effectively organizes these signals into compact and interpretable rules, enabling robust and transparent relation prediction.

Bridging Neural Attribution and Symbolic Features To examine the neuro-symbolic bridge in the proposed framework, we analyzed representative validation examples using two complementary views: (i) node-level attribution on the R-GCN encoder through occlusion-based importance, and (ii) the activated symbolic feature space used by the Tsetlin Machine (TM). This analysis allows us to study how continuous graph-based representations are transformed into discrete relational evidence.

Across examples as shown in Table 7, the neural attribution is consistently concentrated on the head and tail entity nodes together with nearby lexical roots. This indicates that the R-GCN focuses on a small, relation-bearing subgraph rather than diffusing attention across the full sentence. On the symbolic side, the TM activates compact feature sets encoding entity-type compatibility, graph connectivity, sentence-level co-occurrence, and distance constraints. The correspondence between these two views provides

Feature Category	Feature Examples	Description
Distance-based	DIST_LE_2, DIST_LE_5, DIST_GT_5	Capture token-level proximity between head and tail entities. Used to distinguish local vs. long-range relations.
Entity Position	HEAD_BEFORE_TAIL, TAIL_BEFORE_HEAD	Represent relative ordering of entities in a sentence, useful for handling flexible word order in Manipuri.
Structural (Graph-based)	SAME_SENTENCE, GRAPH_CONNECTED	Encode graph connectivity and contextual co-occurrence within the constructed sentence graph.
Entity-Type & Pair-Type	HEAD_TYPE_*, TAIL_TYPE_*, PAIR_*	* represent entity categories (e.g., PER, ORG, LOC) and valid pair combinations (e.g., PAIR_PER_TITLE).
Morphology-aware	HAS_GENITIVE, HAS_LOCATIVE, HEAD_HAS_*, BETWEEN_HAS_*	Capture suffix-level grammatical markers and affix-based linguistic roles in Manipuri.
Embedding-based	EMB_HIGH_k, EMB_POS_k	Binarized R-GCN embedding features providing fine-grained contextual signals for refining symbolic decisions.

Table 6. Feature categories used by the Tsetlin Machine for relation classification

Table 7. Neuro-symbolic bridge analysis with morphological evidence.

Example	Gold	Neural	Top Neural Nodes	Morphological Evidence	Symbolic Features (TM)
E1: Manipur → India	PART_OF	ASSOCI-ATED_WITH	India [ENT], Manipur [ROOT]	Genitive marker (-gi) indicating pos-session/part-whole	HEAD_TYPE_LOC, PAIR_LOC_LOC, GRAPH_CONNECTED
E2: Injury → Hospi-tal	LO-CATED_IN	LO-CATED_IN	Injury [ROOT], Hos-pital [ENT]	Locative suffix (-da) marking loca-tion	PAIR_INJURY_FAC, TAIL_TYPE_FAC, DIST_LE_2
E3: Moreh → Imphal	PART_OF	PART_OF	Moreh [ENT], Im-phal [ENT]	Locative/genitive context (implicit hierarchy)	PAIR_LOC_LOC, HEAD_TYPE_LOC, SAME_SENTENCE
E4: La-horijan → Assam	PART_OF	LO-CATED_IN	Assam [ENT], La-horijan [ENT]	Locative suffix (-da) supporting spatial containment	PAIR_LOC_LOC, GRAPH_CONNECTED
E5: Bomb Attack → Agartala	LO-CATED_IN	LO-CATED_IN	Agartala [ENT], Bomb/Attack [ROOT]	Locative marker (-da) linking event to place	PAIR_EVENT_LOC, HEAD_TYPE_EVENT, TAIL_TYPE_LOC

evidence that the symbolic classifier is grounded in the same local graph structure emphasized by the neural encoder.

A clear example is a LOCATED_IN (E5 of Table 7) instance involving an event and a location. For the pair *bomb attack* → *Agartala*, the neural attribution assigns highest importance to the location entity node and the lexical roots corresponding to the event trigger. The TM prediction for the same pair activates symbolic literals such as HEAD_TYPE_EVENT, PAIR_EVENT_LOC, GRAPH_CONNECTED, SAME_SENTENCE, and TAIL_TYPE_LOC. This shows a strong alignment between neural and symbolic reasoning: the encoder identifies the event–location region of the graph, while the TM converts that evidence into an interpretable location rule.

A second pattern emerges for place hierarchy relations labeled PART_OF (E3 of Table 7). For examples such as *Moreh* → *Imphal* and *Lahorijan* → *Assam*, the neural attribution is dominated by the two location entity nodes and nearby locative roots. The TM, in turn, activates a stable symbolic pattern consisting of HEAD_TYPE_LOC, TAIL_TYPE_LOC, or TAIL_TYPE_LOC_OR_STATE, together with PAIR_LOC_LOC, GRAPH_CONNECTED, and SAME_SENTENCE. These examples suggest that the TM learns a compact place–place or place–state hierarchy rule over graph-derived features.

Importantly, the bridge analysis also highlights the corrective role of symbolic reasoning. In some PART_OF(E1 of Table 7) cases, the neural relation head predicts broader labels such as ASSOCIATED_WITH or LOCATED_IN, whereas the TM correctly predicts PART_OF. This indicates that the symbolic layer is not merely mirroring the neural output, but imposing a more stable decision boundary using explicit relational constraints. In other words, the encoder provides rich contextual representations, while the TM sharpens those representations into discrete, interpretable relation rules.

Overall, these findings support the central claim of the proposed framework: the R-GCN captures relation-bearing subgraphs in a continuous space, and the TM transforms this information into explicit symbolic evidence. The result is a neuro-symbolic pipeline in which neural attribution and symbolic activation are aligned, improving both interpretability and predictive stability.

End-to-End Performance

To mitigate the dominance of the NONE class, we evaluate the model under a *positive-only relation classification* setting, where only valid entity pairs are considered. This isolates the model's ability to distinguish fine-grained relation types.

Experimental Setting. After candidate pruning, 495 positive relation instances are retained. For each pair, contextual embeddings are obtained from the R-GCN encoder, and 379 binary symbolic features encoding structural, morphological, and semantic cues are extracted for the Tsetlin Machine (TM).

Overall Performance. The proposed framework achieves strong and balanced performance:

- Accuracy: 91.72%
- Micro-F1: 91.72%
- Macro-F1: 91.16%

The close alignment between Micro-F1 and Macro-F1 indicates consistent performance across both frequent and rare classes.

Fine-Grained Results and Feature Analysis. Table 8 reports per-class performance, while Table 9 analyzes the contribution of structural and morphology-aware cues.

Relation	Precision	Recall	F1
AFFECTS	0.97	0.97	0.97
AFFILIATED_WITH	1.00	0.85	0.92
ASSOCIATED_WITH	0.85	0.97	0.90
HAS_VALUE	1.00	0.88	0.93
HOLDS_TITLE	1.00	0.88	0.93
LOCATED_IN	0.92	0.96	0.94
MIXED_WITH	1.00	0.88	0.93
OPPOSES	0.97	0.83	0.89
PARTICIPATES_IN	1.00	0.87	0.93
PART_OF	0.94	0.79	0.86
TREATS	1.00	0.57	0.73
USES	1.00	1.00	1.00

Table 8. Per-class performance

Relation	Morph Cues	Struct Cues	Accuracy
AFFECTS	0/36	36/36	0.97
AFFILIATED_WITH	5/33	33/33	0.79
ASSOCIATED_WITH	1/160	160/160	0.97
HAS_VALUE	0/32	32/32	0.88
HOLDS_TITLE	1/16	16/16	0.94
LOCATED_IN	1/106	106/106	0.96
MIXED_WITH	0/8	8/8	0.88
OPPOSES	7/35	35/35	0.83
PARTICIPATES_IN	0/15	15/15	0.87
PART_OF	1/38	38/38	0.82
TREATS	0/7	7/7	0.57
USES	0/9	9/9	1.00

Table 9. Structural vs. morphological cue distribution

Analysis. Three key observations were noticed:

(1) Structural dominance. Structural cues are present in all instances (100% coverage), forming the primary basis for relation prediction. High-performing relations such as AFFECTS and LOCATED_IN achieve strong results even with minimal morphological support.

(2) Selective role of morphology. Morphological cues appear sparsely but improve precision for specific relations such as OPPOSES and AFFILIATED_WITH, acting as complementary refinements rather than primary signals.

(3) Ambiguity-driven errors. Lower-performing relations (e.g., PART_OF, TREATS) exhibit weak morphological and structural distinctiveness, leading to reduced recall.

Category	Count	Percentage
Total	495	100%
Correct	454	91.72%
Errors	41	8.28%

Table 10. Error distribution

Error Analysis. Errors are sparse (8.28%) and concentrated among semantically overlapping relations. In particular, ASSOCIATED_WITH frequently acts as a fallback class when relation-specific cues are weak. Confusions are most common among PART_OF, OPPOSES, and TREATS, reflecting shared morphological realizations (e.g., genitive markers).

Rare classes such as TREATS suffer from data sparsity, limiting reliable rule formation. In contrast, frequent relations benefit from repeated structural patterns, leading to stable predictions.

Summary The results demonstrate that structural graph signals form the backbone of relation extraction, while morphology-aware cues provide targeted refinements. The integration of R-GCN representations with TM-based symbolic reasoning enables accurate, interpretable, and robust performance in low-resource settings.

Baseline Comparison

Table 12 compares the proposed framework against representative sequential, graph-based, and transformer baselines.

The proposed R-GCN + TM model outperforms all baselines across every relation type, with especially large gains for structurally and morphologically complex relations such as HOLDS_TITLE, AFFILIATED_WITH, and TREATS. Sequential models perform moderately on frequent surface-level relations but struggle with free word order and affix-driven grammatical roles. The GCN baseline improves on structurally grounded relations such as LOCATED_IN and PARTICIPATES_IN, confirming the importance of graph structure, but remains insufficient for morphology-sensitive cases. XLM-R performs poorly on nearly all classes and tends to collapse toward broader labels, showing that multilingual transfer alone is ineffective without explicit linguistic grounding.

Overall, the baselines fail because they do not jointly model *morphology, graph structure, and symbolic constraints*. Sequential models are limited by token order and local context (Huang et al. 2015; Zeng 2014); dependency-based GCNs rely on noisy or unavailable parses and do not explicitly encode affixal roles (Zhang

Table 11. Confusion matrix for relation classification

Gold \ Pred	AFF	AFFL	ASSOC	HAS_V	HOLD	LOC	MIX	OPP	PART_IN	PART_OF	TREAT	USE
AFFECTS	**35**	0	1	0	0	0	0	0	0	0	0	0
AFFILIATED_WITH	0	**28**	5	0	0	0	0	0	0	0	0	0
ASSOCIATED_WITH	1	0	**155**	0	0	4	0	0	0	0	0	0
HAS_VALUE	0	0	3	**28**	0	0	0	0	0	1	0	0
HOLDS_TITLE	0	0	1	0	**14**	1	0	0	0	0	0	0
LOCATED_IN	0	0	2	0	0	**102**	0	1	0	1	0	0
MIXED_WITH	0	0	0	0	0	1	**7**	0	0	0	0	0
OPPOSES	0	0	5	0	0	1	0	**29**	0	0	0	0
PARTICIPATES_IN	0	0	2	0	0	0	0	0	**13**	0	0	0
PART_OF	0	0	7	0	0	1	0	0	0	**30**	0	0
TREATS	0	0	2	0	0	1	0	0	0	0	**4**	0
USES	0	0	0	0	0	0	0	0	0	0	0	**9**

Relation	BiLSTM	CNN–BiLSTM	GCN	XLM-R	R-GCN + TM
AFFECTS	0.67	0.73	0.67	0.00	**0.97**
AFFILIATED_WITH	0.00	0.21	0.21	0.00	**0.92**
ASSOCIATED_WITH	0.57	0.56	0.63	0.44	**0.90**
HAS_VALUE	0.57	0.45	0.37	0.00	**0.93**
HOLDS_TITLE	0.00	0.00	0.25	0.00	**0.93**
LOCATED_IN	0.48	0.55	0.65	0.14	**0.94**
MIXED_WITH	0.00	0.00	0.67	0.00	**0.93**
OPPOSES	0.12	0.00	0.40	0.00	**0.89**
PARTICIPATES_IN	0.80	0.80	1.00	0.00	**0.93**
PART_OF	0.27	0.38	0.36	0.00	**0.86**
TREATS	0.00	0.00	0.00	0.00	**0.73**
USES	0.67	0.67	0.50	0.00	**1.00**

Table 12. Fine-grained F1 comparison across baseline models and the proposed framework

Model Variant	Accuracy	Micro-F1	Macro-F1
Full Neural (R-GCN + MLP)	0.27	0.27	0.16
No Graph Structure	0.35	0.35	0.31
No Morphology	0.36	0.36	0.17
Proposed (R-GCN + TM)	**0.9172**	**0.9172**	**0.9116**

Table 13. Ablation study evaluating the contribution of graph structure, morphology, and symbolic reasoning

et al. 2018); and multilingual transformers such as XLM-R lack sufficient exposure to Manipuri-specific morphological patterns (Conneau 2020). In contrast, the proposed framework combines morphology-aware graph encoding with interpretable symbolic reasoning, enabling more robust and balanced relation extraction under limited supervision.

Ablation Analysis

Table 13 summarizes the contribution of graph structure, morphology-aware features, and symbolic reasoning.

Three conclusions follow. First, replacing the Tsetlin Machine with a neural classifier causes the largest drop in performance, reducing Macro-F1 from 0.91 to 0.16. This suggests that the graph embeddings alone are not sufficient unless they are coupled with a decision mechanism that can exploit sparse, high-value relational cues. Neural classifiers are known to be vulnerable to overfitting and instability under data sparsity and class imbalance (Goodfellow et al. 2016; Wang et al. 2021).

Second, removing graph structure substantially degrades performance (Macro-F1 = 0.31), confirming that explicit relational connectivity is essential for capturing long-range and non-linear interactions between entities (Zhang et al. 2018; Schlichtkrull et al. 2018b). Third, removing morphology-aware features reduces Macro-F1 to 0.17, showing that suffix-level grammatical information remains critical for distinguishing relations in an agglutinative language (Sennrich et al. 2016; Bolucu et al. 2021).

Role of symbolic reasoning. The strongest gain comes from symbolic reasoning. Unlike neural classifiers, the TM operates on discrete graph-derived features and learns propositional rules through reinforcement-driven updates (Granmo 2018; Abeyrathna et al. 2021). This makes it well suited to low-resource settings, where informative signals such as entity-type compatibility, graph connectivity, and morphological markers are sparse but highly discriminative. By directly encoding these signals as logical clauses, the TM provides both stable decision boundaries and interpretable reasoning. The ablation results therefore confirm that high performance emerges only when morphology-aware graph encoding is combined with symbolic classification.

Data Availability

The dataset and code used in this study are publicly available to support reproducibility and further research. The Manipuri Entity Relationship Extraction Dataset (MERED), along with the complete implementation of the proposed R-GCN + Tsetlin Machine framework and baseline models, can be accessed at:

https://github.com/jimmy3018/manipuri-relationship-extraction-Graph-GCN-TM-

The repository includes:

- The MERED dataset with annotated entity pairs and relation labels,
- Preprocessing scripts for morphology-aware graph construction,
- Implementation of the R-GCN encoder and Tsetlin Machine classifier,
- Baseline models (BiLSTM, CNN–BiLSTM, GCN, and XLM-R),
- Evaluation scripts and configuration files for reproducing experimental results.

All experiments reported in this paper can be reproduced using the provided code and dataset. Any updates or extensions to the dataset and models will be maintained through the same repository.

Conclusion

This work presents a neuro-symbolic framework for relation extraction in low-resource, morphologically rich languages, with a focus on Manipuri. The proposed approach constructs heterogeneous graphs from morphologically segmented text, where both lexical roots and grammatical affixes are modeled as nodes. Contextual representations are learned using a relational graph convolutional network (R-GCN), and a Tsetlin Machine (TM) performs interpretable classification over binary symbolic features derived from graph structure, entity types, and binarized embeddings.

Experimental results on the MERED dataset demonstrate the effectiveness of the approach. Under a positive-only classification setting, the proposed model achieves an accuracy of **91.72%**, with **Micro-F1 of 91.72%** and **Macro-F1 of 91.16%**, indicating consistent performance across both frequent and low-frequency relation classes. Fine-grained evaluation shows strong performance for structurally grounded relations such as `AFFECTS` (0.97), `LOCATED_IN` (0.94), and `HOLDS_TITLE` (0.93). Furthermore, the model significantly outperforms all baselines across every relation type, including challenging morphology-driven relations such as `AFFILIATED_WITH` and `TREATS`, where conventional neural models fail.

The results reveal that structural graph signals form the primary backbone of relation prediction, while morphology-aware features provide selective, high-precision refinements. The inclusion of symbolic reasoning through the Tsetlin Machine enables the model to learn explicit logical clauses, improving robustness under data sparsity and class imbalance while maintaining interpretability.

Beyond performance gains, the framework establishes a transparent bridge between neural and symbolic learning. By transforming continuous graph embeddings into discrete logical rules, the model provides insight into how linguistic structure is encoded and utilized for relation prediction. This makes the approach particularly suitable for low-resource settings where both accuracy and interpretability are critical.

Limitations

Despite its effectiveness, the proposed framework has several limitations.

First, the dataset size remains relatively small, which restricts the diversity of linguistic patterns available for learning. While the TM mitigates data sparsity to some extent, rare relations such as `TREATS` still suffer from lower recall due to limited training examples.

Second, morphology-aware features are currently incorporated through manually designed or heuristically derived cues. Although effective, this approach may not capture the full richness of morphological variation, particularly in cases involving complex affix stacking or ambiguous grammatical roles.

Third, the framework relies on candidate pair generation and filtering, which may introduce bias by excluding difficult or ambiguous cases. Errors in earlier pipeline stages can propagate to the final prediction, limiting overall performance.

Fourth, while the TM provides interpretability, a large number of clauses can make global interpretation challenging. Although individual rules are understandable, summarizing the full decision space remains non-trivial.

Future Work

Future work will focus on extending the framework along several directions.

First, we plan to expand the MERED dataset with more diverse domains and larger annotation coverage, enabling more robust evaluation and improved generalization. Incorporating semi-supervised or weakly supervised data augmentation techniques may further alleviate data sparsity.

Second, we aim to improve morphology modeling by integrating neural morphological analyzers that can better capture complex affix interactions. This could enhance the representation of grammatical roles in the graph.

Third, we intend to strengthen the neuro-symbolic bridge by aligning neural attribution methods (e.g., GNNExplainer) with TM clause activations, enabling more precise mapping between subgraph evidence and symbolic rules.

Fourth, extending the framework to cross-lingual and multilingual settings is a key direction. Since the approach is not dependent on syntactic parsers, it can potentially be applied to other agglutinative and low-resource languages with minimal adaptation.

Overall, this work demonstrates that combining morphology-aware graph representations with symbolic reasoning provides a promising direction for interpretable and robust relation extraction in low-resource language settings.

References

K. D. Abeyrathna et al. A survey of the tsetlin machine. *IEEE Access*, 2021.

T. Bach et al. A web of semantic data. *IEEE Intelligent Systems*, 2004.

P. Basumatary. Morphological analysis of bodo language. *IJLT*, 2014.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

B. Bickel. The role of morphology in argument marking. *Annual Review of Linguistics*, 1:85–104, 2015.

A. Bisazza and M. Federico. A survey of word reordering in statistical machine translation. *Computational Linguistics*, 40(2):399–447, 2014.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. In *TACL*, 2017.

N. Bolucu, O. A. Can, and G. Eryiğit. Semantically enhanced graph neural networks for turkish named entity recognition. *Pattern Recognition Letters*, 147: 98–104, 2021.

B. Comrie. *Language Universals and Linguistic Typology*. University of Chicago Press, 1989a.

B. Comrie. *Language Universals and Linguistic Typology*. University of Chicago Press, 1989b.

A. e. a. Conneau. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020.

R. Cotterell et al. On the role of morphological information in multilingual nlp. In *NAACL*, 2018a.

R. Cotterell et al. A survey of morphological learning in nlp. *Transactions of the ACL*, 6:1–20, 2018b.

M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), 2007. doi: 10.1145/1187415.1187418.

S. Debbarma. Morphological structure of kokborok. *Indian Linguistics*, 2017.

N. C. Devi and T. D. Singh. Challenges in processing manipuri language. *Language Resources and Evaluation*, 53(2):285–314, 2019. doi: 10.1007/s10579-018-9445-7.

J. e. a. Devlin. Bert: Pre-training of deep bidirectional transformers. In *NAACL*, 2019.

A. d'Avila Garcez et al. Neuro-symbolic ai: The state of the art. *arXiv*, 2019.

S. Garg and R. Sharma. Morphological segmentation improves nlp for indian languages. *Knowledge-Based Systems*, 238:107832, 2022. doi: 10.1016/j.knosys.2021.107832.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

O.-C. Granmo. The tsetlin machine—a game theoretic bandit driven approach to optimal pattern recognition with propositional logic. *arXiv preprint arXiv:1804.01508*, 2018.

M. Haspelmath. *Understanding Morphology*. Arnold, 2002.

M. Haspelmath. Understanding morphology. *Language*, 2010.

J. He, Y. Wang, and X. Liu. Graph neural networks for morphology-aware relation extraction in agglutinative languages. *Knowledge-Based Systems*, 275:110746, 2023. doi: 10.1016/j.knosys.2023.110746.

X. e. a. He. A graph neural network approach for relation extraction in uyghur. *Applied Intelligence*, 2023.

P.-Y. Huang, Y. Shen, J. Liu, J. Gao, and W. Chen. Neural machine translation of pro-drop languages with pronoun prediction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1038–1048, 2018. doi: 10.18653/v1/D18-1127.

Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

B. Ji, T. Liu, X. Zhang, and S. Li. A graph-based approach for joint named entity recognition and relation extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2078–2091, 2022. doi: 10.1109/TASLP.2022.3174567.

J. Kanerva and F. Ginter. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of CoNLL 2018*, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

P. Lahoti, N. Mittal, and G. Singh. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2022. doi: 10.1145/3548457.

Q. Li and H. Ji. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *COLING*, 2016.

Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11604.

Y. Ma, R. Zhang, and W. Li. Limitations of transformer models in low-resource morphologically rich languages. *Expert Systems with Applications*, 195:116552, 2022. doi: 10.1016/j.eswa.2022.116552.

M. Miwa and M. Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1105–1116, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1105. URL https://aclanthology.org/P16-1105/.

S. Murmu. A study of santali morphology. *Language in India*, 2015.

J. e. a. Nivre. Universal dependencies v1. *LREC*, 2016.

J. Philip, S. Siripragada, V. P. Namboodiri, and C. V. Jawahar. Revisiting low resource status of indian languages in machine translation. 2020. doi: 10.48550/arXiv.2008.04860.

A. Rahimi, Y. Li, and T. Cohn. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 151–164, Florence, Italy, 2019. Association

for Computational Linguistics. doi: 10.18653/v1/P19-1015.

M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *Proceedings of ESWC*, 2018a. doi: 10.1007/978-3-319-93417-4_38.

M. Schlichtkrull et al. Modeling relational data with graph convolutional networks. In *ESWC*, 2018b.

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016. doi: 10.18653/v1/P16-1162.

A. K. Singh and D. M. Sharma. Morphological segmentation for dravidian languages. *Journal of South Asian Languages and Linguistics*, 8(2):165–189, 2021. doi: 10.1515/jsall-2021-0007.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

R. Tsarfaty et al. Morphologically rich languages in nlp. *Computational Linguistics*, 2010a.

R. Tsarfaty et al. Statistical parsing of morphologically rich languages: What, how and whither. *Proceedings of NAACL Workshop*, 2010b.

B. Wang et al. A survey on low-resource natural language processing. *ACM Computing Surveys*, 2021.

Y. Wang et al. Ace05 relation extraction with neural models. In *ACL*, 2020.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021a. doi: 10.1109/TNNLS.2020.2978386.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021b.

L. Yao et al. Knowledge graph embeddings: A survey. *IEEE TKDE*, 2019.

D. Zeman, M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, et al. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task*, 2017.

D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1753–1762. Association for Computational Linguistics, 2015.

D. e. a. Zeng. Relation classification via cnn. In *COLING*, 2014.

X. Zeng, D. Zeng, S. He, K. Liu, and J. Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 506–514, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1047. URL https://aclanthology.org/P18-1047/.

Y. Zhang, P. Qi, and C. D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, 2018. doi: 10.18653/v1/D18-1244.

X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, and R. Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. 2023. doi: 10.48550/arXiv.2306.02051.

J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. doi: 10.1016/j.aiopen.2021.01.001.

H. Zhu, Y. Lin, Z. Liu, J. Fu, T.-S. Chua, and M. Sun. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 593–603, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1128. URL https://aclanthology.org/P19-1128/.