Language Understanding $\lambda(\mu)$ Laboratory

myNER: Contextualized Burmese Named Entity Recognition with Bidirectional LSTM and fastText Embeddings via Joint Training with POS Tagging

Kaung Lwin Thant, Kwankamol Nongpong, Ye Kyaw Thu

Thura Aung, Khaing Hsu Wai, Thazin Myint Oo

Presented at ICCI 2025, Chonburi, Thailand

Outline of Presentation

- Introduction
- Corpus Development
- Methodology Overview
- Experiment Setting
- Results
- Limitations
- Future Works

Introduction

What is NER?

- Identifies names, dates, orgs, locations
- Extracts structure from unstructured text
- Key NLP task for information extraction
- Supports MT, QA, search, chatbots
- Requires accurate, context-aware tagging
- Modeled as token-level classification
- Needs annotated corpora for training

Challenges in Burmese NER

- Burmese is low-resource and underrepresented
- Public annotated corpora are very limited
- Complex script and rich morphology
- No word boundaries (no whitespace)
- Non-Latin script complicates preprocessing
- Domain shifts affect entity tag consistency
- Manual annotation is labor-intensive

Our Contributions

- Word-level NER corpus named myNER_7tags
- BIOES scheme + 7 NER tag categories
- POS tags added for richer syntax signals
- Corpus has 16,605 manually tagged sentences
- Traditional CRF and NN-based BiLSTM-CRF models evaluated
- Achieved up to 98% overall accuracy, SOTA results

Tag	Description	Examples
DATE	Date	၁ ရက် ၅ လ ၁၉၉၆ (01-05-1996), ကဆုန်လပြည့် (full moon day of Kasone)
TIME	Time	မနက် ၆ နာရီ ၁၀ မိနစ် (06:10 AM)
NUM	Numbers	၁၀၀ (100), ၁ သန်း (1 Million)
PER	Person Names	အန်တိုနီယို (Antonio), ကောင်းလွင်သန့် (Kaung Lwin Thant)
ORG	Organizations and Instiutions	မိုက်ခရိုဆော့ (Microsoft Inc.), ဟားဗတ်တက္ကသိုလ် (Harvard University)
LOC	Locations and Geographic Features	အာရှတိုက် (Asia), ထိုင်းနိုင်ငံ (Thailand)
O	Outside	Non-specific entities which do not match the above tags

Table 1: Tag Names, Description, and Examples of myNER Corpus

Corpus Development

- Built mostly upon the myPOS Corpus (version 3):Linguistically diverse and representative Burmese text.
- Contains 16,605 manually annotated sentences, suitable for academic NLP experiments.

Corpus Development (Cont'd)

- Token distribution shown in Fig. 1
- 13,000+ location names
- 5,000+ numbers (NUM)
- 3,000+ personal names (PER)
- 600+ organizations (ORG)
- Tag imbalances are still challenging factor

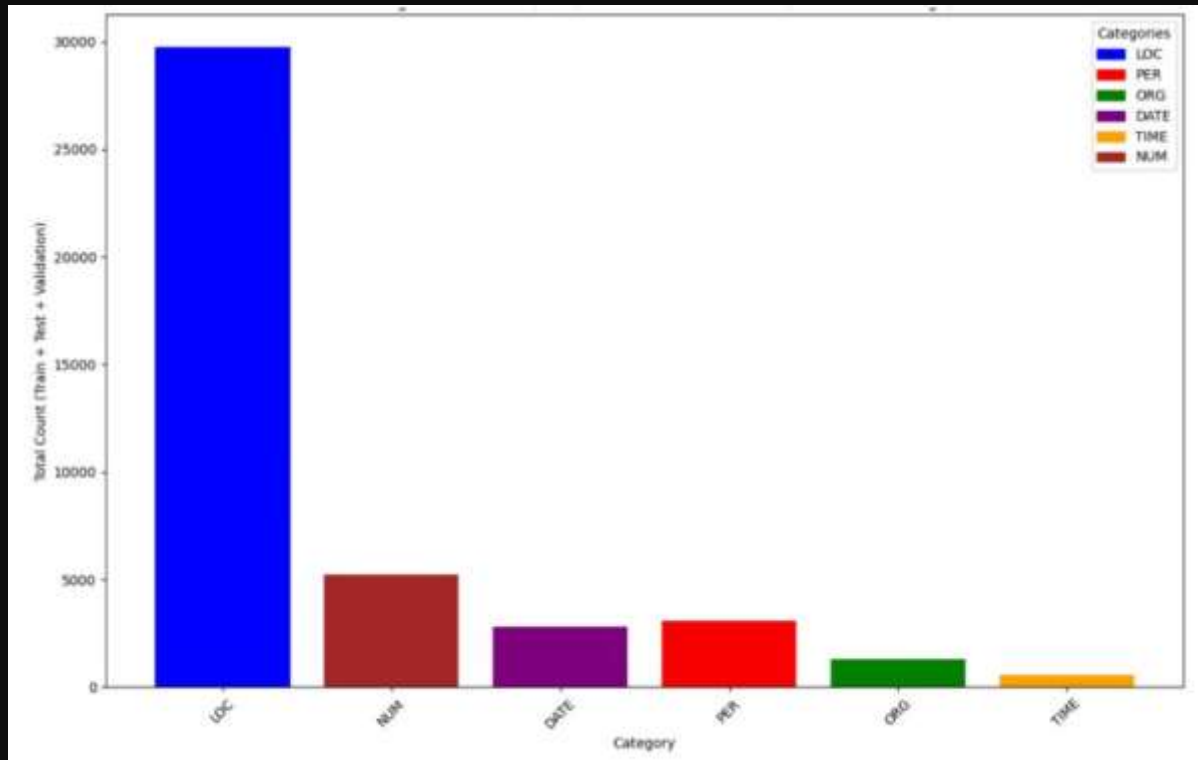

Fig. 1: Tag categories except main O Tag in the dataset

Corpus Development (Cont'd)

- BIOES scheme marks entity boundaries
- S-LOC = Single-token location
- B-ORG/E-ORG = Multi-token organization
- POS tags provide syntactic context
- O = Non-entity tokens

Word	POS Tag	NER Tag
မန္တလေး	n	S-LOC
ဌာ်	ppm	O
ရန်ကုန်	n	B-ORG
တက္ကသိုလ်	n	E-ORG
လက်အောက်ခံ	n	O
ဆေးအတတ်သင်	n	B-ORG
ကောလိပ်	n	E-ORG
ခွဲရိုး	v	O
ခဲ့	part	O
သည်	ppm	O

TABLE 2: myNER sample data with CoNLL format

Methodology Overview

- Input: Pretrained fastText embeddings
- Context: BiLSTM (forward + backward)
- Output: CRF/Softmax inference
- Key: Joint POS+NER training

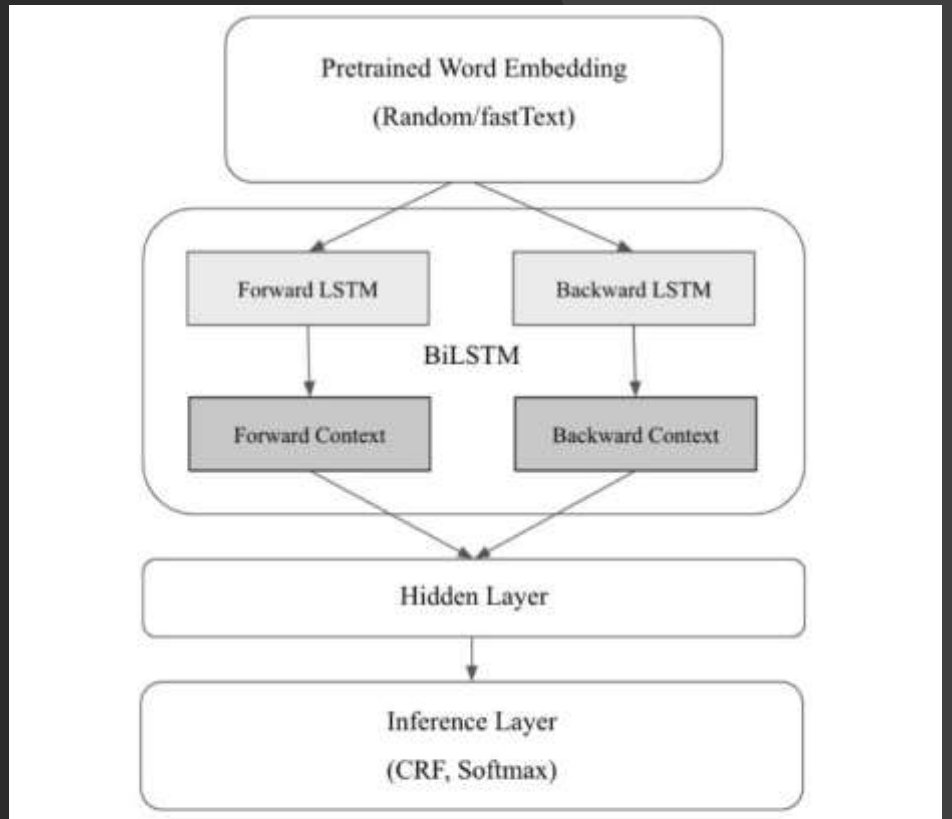

Fig. 2: Word Embeddings + BiLSTM Sequence Tagging

Joint Modeling for POS and NER Approach

- Embedding Layer: fastText word vectors/random embedding
- BiLSTM Layer captures sentence context
- Shared Layer: used by both POS and NER
- Trained together with shared BiLSTM encoder
- Output Layer: two CRFs (one per task)
- POS predicts grammar tags (noun, verb, etc.)
- NER predicts entity types (PER, LOC, ORG)

fastText Embeddings

- Subword-aware (handles OOV/rare words)
- Morphology semantics in one model
- 300-dim vectors (Burmese Wikipedia)
- Works in fixed or fine-tuned mode
- Fine-tuning boosts task accuracy
- Good fit for Burmese morphology

Traditional CRF Model with Feature Engineering

- CRFs are probabilistic models used for structured prediction, well-suited for sequence labeling tasks like POS and NER.
- **Feature Engineering**
  - Lexical Features:
    - Current, previous, and next words
    - First/last word indicators
  - Morphological Features:
    - Prefixes & suffixes (1–3 chars)
    - Hyphenation, numeric detection (supports DATE/NUM recognition)
  - fastText-enhanced:
    - Embedding vector average included as a dense feature
    - Adds semantic generalization to statistical structure

BiLSTM + CRF Architecture

- **BiLSTM-CRF** architecture integrates contextual encoding with structured output prediction.
  - **BiLSTM Encoder**: Encodes contextual representations for each token.
  - **CRF Layer**: Models label transitions to ensure coherent tag sequences (avoids invalid transitions like I-LOC following B-PER).
  - Combines deep contextual understanding (BiLSTM) with global sequence optimization (CRF).
  - Significantly improves accuracy in tasks where label dependencies matter.

Experiment Setting

- **Training Environment**
  - Platform: Kaggle
  - GPU: NVIDIA Tesla P100 (16 GB
- **Task Configurations**
  - Single-task (NER)
  - Joint-task (NER + POS tagging)
- **Dataset Split**
  - Split: 80% Train – 10% Validation – 10% Test

Experiment setting(Hyperparameters)

CRF (Traditional ML)

- Library: CRFsuite
- Embedding: Pretrained fastText (300-dim)
- Feature set:
  - Prefixes/Suffixes (1–3 chars)
  - Word shape, numeric/hyphen flags
  - Position in sentence (first/last)
  - Context window (previous & next word)

Setting	Random Embedding	fastText (Fixed)	fastText (Fine-tuned)
Embedding Dim	300	300	300
BiLSTM Hidden Units	128	256	256
Batch Size	32	64	64
Dropout	0.5	0.5	0.5
Optimizer	Adam (lr=0.001, β_1=0.9, β_2=0.999)	Same	Same
Epochs	Max 50 + Early Stopping (on val loss)	Same	Same
Inference Layer	Softmax or CRF	Softmax or CRF	Softmax or CRF

Table 3: Hyperparameters of best-performing BiLSTM-based models

BiLSTM-Based Model Hyperparameters

- All models were implemented using the PyTorch deep learning framework.
- Tuned hyperparameters for best model performance
- Tested multiple embedding, hidden, batch, dropout configs

Results

The table compares NER performance across models using different training strategies, embeddings, and inference layers.

Three metrics were used: Accuracy, Weighted F1, and Macro F1

Model	Embeddings	Training	Accuracy	F1 (Weighted)	F1 (Macro)
CRF	w/o fastText	Single (NER)	**0.9818**	**0.9812**	0.7405
		Joint (POS+NER)	0.9812	0.9807	0.7367
	with fastText	Single (NER)	**0.9818**	0.9811	**0.7429**
		Joint (POS+NER)	0.9810	0.9804	0.7345
BiLSTM-Softmax	Random	Single (NER)	0.9740	0.9725	0.6478
		Joint (POS+NER)	0.9730	0.9714	0.6463
	fastText (Frozen)	Single (NER)	0.9737	0.9723	0.6578
		Joint (POS+NER)	0.9753	0.9734	0.6489
	fastText (Fine-tuned)	Single (NER)	0.9783	0.9779	0.6502
		Joint (POS+NER)	0.9780	0.9764	0.6743
BiLSTM-CRF	Random	Single (NER)	0.9740	0.9730	0.6784
		Joint (POS+NER)	0.9742	0.9730	0.6907
	fastText (Frozen)	Single (NER)	0.9747	0.9729	0.6396
		Joint (POS+NER)	0.9746	0.9737	0.6790
	fastText (Fine-tuned)	Single (NER)	0.9755	0.9753	0.7154
		Joint (POS+NER)	**0.9791**	**0.9776**	**0.7395**

Table 4: Performance Comparison of Single NER and Joint POS+NER Models

Tag	B	I	E	S
LOC	0.9763	0.9711	0.9766	0.6849
DATE	0.8281	0.7654	0.8682	0.9136
NUM	0.4211	–	0.4211	0.9501
PER	0.9143	–	0.8986	0.5783
ORG	0.6265	0.5217	0.5977	0.5882
TIME	0.8889	–	0.8889	–
O	0.9897	–	–	–

Table 5:Tag-wise F1-Scores for Joint BiLSTM-CRF Model with Fine-Tuned fastText Embeddings

Tag-wise Evaluation Results

- Model captures LOC and DATE well with high F1.
- Still promising performance on sparse tags like NUM and ORG despite imbalance.

Result Discussion

- Best Macro F1: 0.7429 (CRF + fastText, single-task)
- Best Weighted F1: 0.9811 — state-of-the-art result
- Best Joint Model: BiLSTM-CRF + fine-tuned fastText
- Macro F1 (joint model): 0.7395 — class-balanced performance
- High-performing tags: LOC, DATE — due to higher frequency
- Moderately performing tags: ORG, NUM — impacted by data sparsity
- Joint training improves generalization with POS sharing

Limitations

- **Dataset Size:** The dataset used was relatively limited in scale, which may restrict generalization across diverse linguistic patterns in Burmese.
- **Class Imbalance:** A high dominance of the "O" tag and sparse samples for entities like TIME, ORG, and NUM affected macro-level performance.
- **Low-Frequency Tags:** Tags with fewer training examples, such as NUM and TIME, led to noticeably lower macro F1 scores due to insufficient learning.
- **Handling OOV & Domain-Specific Terms:** Despite fastText's subword capability, the models still struggle with rare or domain-specific out-of-vocabulary words.
- **Training Efficiency of CRF:** CRF-based models, though more accurate, require significantly higher training time (e.g., 257s vs. 34s for Softmax in joint setting), which limits scalability.

Future Works

- **Enhance Dataset Quality & Balance:** Expand the corpus with more annotated data across diverse domains and apply class balancing techniques to improve performance on low-frequency tags (e.g., NUM, TIME, ORG).
- **Leverage Pretrained Language Models:** Fine-tune large multilingual or Burmese-specific LLMs (e.g., mBERT, XLM-R) and explore transformer-based architectures for richer contextual understanding.
- **Integrate Linguistic Knowledge**: Incorporate additional linguistic features such as syntactic dependencies, morphological patterns, or chunking to support learning.
- **Improve Model Efficiency:** Explore lighter or quantized models for faster inference and real-world deployment in low-resource environments.

Thank You & Questions

- I deeply appreciate your time and attention.

Contact Info:

- Name: Kaung Lwin Thant (Matt)
- Personal Email: kaunglwinthant@gmail.com

Reference

- Hsu Myat Mo and Khin Mar Soe, "Named Entity Recognition for Myanmar Language," in Proceedings of the 2022 International Conference on Communication and Computer Research (ICCR 2022), Sookmyung Women's University, Seoul, Korea, 2022.
- Rrubaa Panchendrarajan and Aravindh Amaresan, "Bidirectional LSTM-CRF for Named Entity Recognition," in Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong, Dec. 2018. Available online
- Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul, "Improving neural machine translation with POS-tag features for low-resource language pairs," Heliyon, vol. 8, no. 8, p. e10375, 2022. DOI
- Ye Kyaw Thu, Thura Aung, Thepchai Supnithi, "Neural Sequence Labeling Based Sentence Segmentation for Myanmar Language," in The 12th Conference on Information Technology and Its Applications (CITA 2023), Lecture Notes in Networks and Systems, vol. 734, Springer. DOIP.
- Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," arXiv preprint arXiv:1607.04606, 2016. arXiv link
- Hsu Myat Mo, Khin Thandar Nwet, and Khin Mar Soe, "CRF-Based Named Entity Recognition for Myanmar Language," in Genetic and Evolutionary Computing (ICGEC 2016), Advances in Intelligent Systems and Computing, vol. 536, Springer, 2017.
- J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the 18th International Conference on Machine Learning, 2001.
- J. Yang, S. Liang, and Y. Zhang, "Design challenges and misconceptions in neural sequence labeling," 2018.