

iSAI-NLP 2024

Nov 11-15, 2024, Pattaya City, Thailand

myOCR: Optical Character Recognition for Myanmar language with Post-OCR Error Correction

Thura Aung*, Ye Kyaw Thu, Myat Noe Oo

Language
Understanding
 $\lambda(\mu)$ Laboratory

NECTEC
a member of NSTDA



* Majority of the work done during Internship at LU Lab



OUTLINES

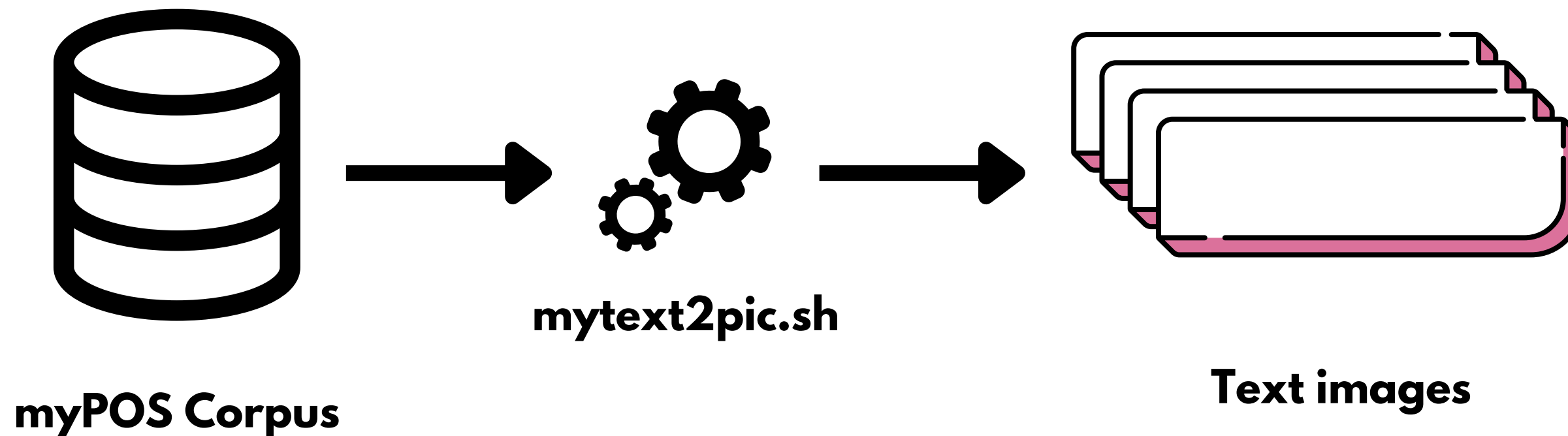
1. Motivation
2. Image Dataset Preparation
3. Methodologies
4. Experiments
5. Conclusion
6. Future Work



Motivation

- **OCR's Role:** Converts physical documents to digital, enabling machine readability
- **Challenges:** Requires accuracy for diverse layouts and graphical info
- **Focus on Graphical Info** - synthetic text image dataset with 14 different font styles
- **OCR Approach**
 - Convolutional Neural Network (CNN) improves OCR accuracy
 - Bi-directional LSTM (BiLSTM-RNN) enhances sequence recognition
 - CRNN + CTC boosts recognition performance
- **Post OCR Challenges:** Diverse fonts & scenarios impact OCR accuracy

Image Dataset Preparation



- **Data Source:** myPOS corpus v3.0, originally for Myanmar POS tagging
- **Font Diversity:** Generated text images using 14 font styles to reduce bias

Image Dataset Preparation (Cont'd)

TABLE I: myOCR IMAGE DATASET PARTITIONING
FOR THE EXPERIMENTS

	Images	Words	Characters
Train	18,052	30,008	1,884,829
Valid	5,802	9,475	599,906
Test	1,936	15,975	201,764
Total	25,790	55,458	2,686,499

Image Dataset Preparation (Cont'd)

TABLE II: SYNTHETIC IMAGES OF နည်းပညာကဏ္ဍ (“TECHNOLOGY SECTOR” IN ENGLISH) WITH DIFFERENT FONTS

Font Name	Text Image	Font Name	Text Image
Burmese Handwriting Style 04	နည်းပညာကဏ္ဍ	Kamjing	နည်းပညာကဏ္ဍ
Myanmar Ayar3	နည်းပညာကဏ္ဍ	Z01-UMoe	နည်းပညာကဏ္ဍ
Z03-Press	နည်းပညာကဏ္ဍ	Z09-LatYaySat	နည်းပညာကဏ္ဍ
Masterpiece Spring Revolution	နည်းပညာကဏ္ဍ	Masterpiece Uni Type	နည်းပညာကဏ္ဍ
Myanmar Chatulight	နည်းပညာကဏ္ဍ	Myanmar Phiskel	နည်းပညာကဏ္ဍ
Myanmar Sanpya	နည်းပညာကဏ္ဍ	Myanmar Yin Mar	နည်းပညာကဏ္ဍ
NKSSmart3	နည်းပညာကဏ္ဍ	Pyidaungsu	နည်းပညာကဏ္ဍ

Image Dataset Preparation (Cont'd)

Number of Lines

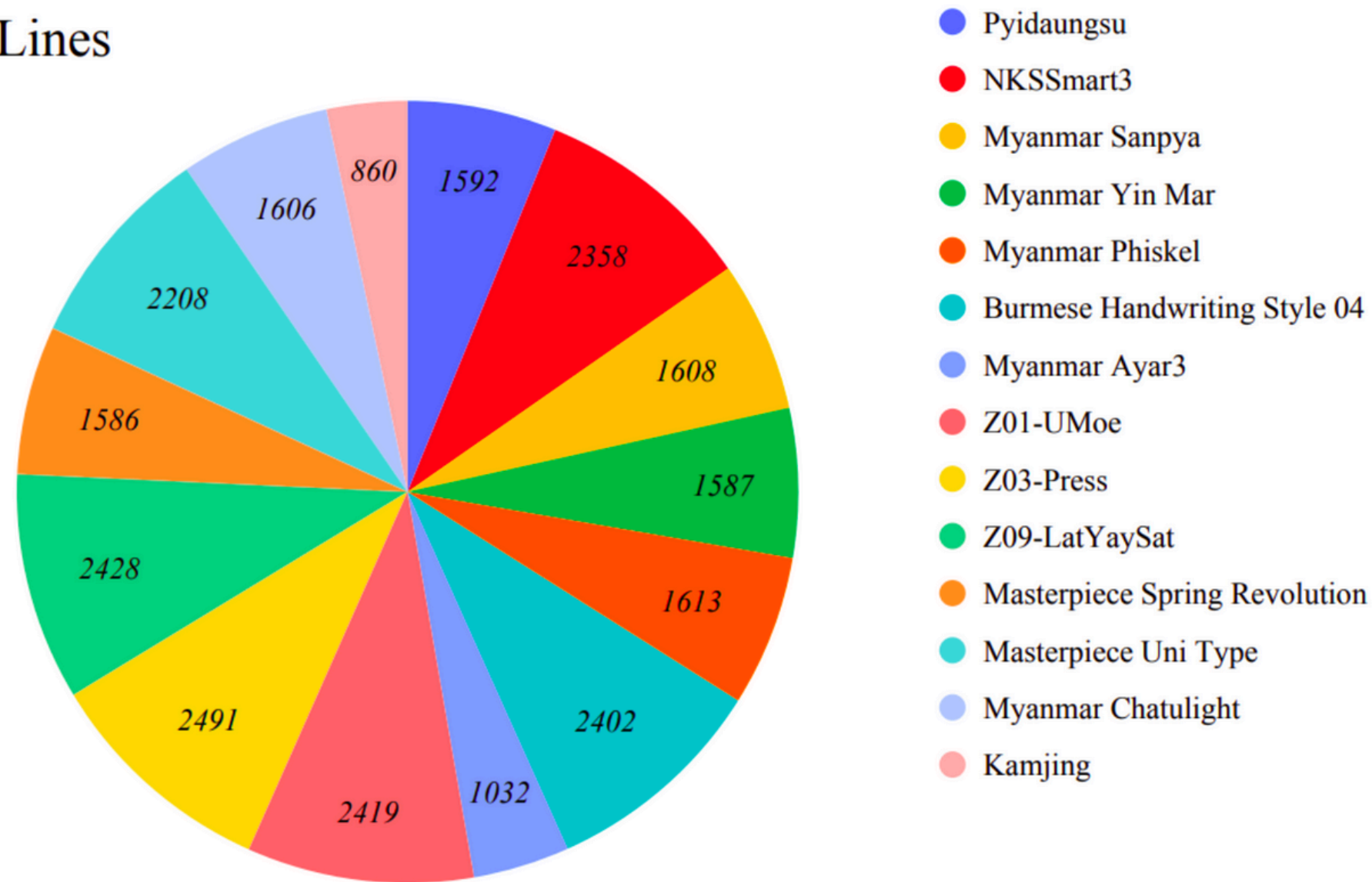


Fig. 1: Number of lines of each font used in the synthetic text images



Methodologies

Optical Character Recognition (OCR)

1. Feature Extraction

- a. Converts grayscale image to feature vector using CNN (VGG architecture)
- b. Focuses on character recognition, ignores color, size, background, etc.

2. Feature Sequence Modeling

Uses BiLSTM to capture bidirectional feature sequences

3. Sequence Decoding

Predicts character sequences with Connectionist Temporal Classification (CTC)



Methodologies (Cont'd)

Post-OCR Correction

Statistical Approaches

- N-gram Similarity:
 - Character-level N-gram similarity for detecting & correcting misspellings
- SymSpell Algorithm:
 - Uses delete-only edit generation for faster candidate lookup efficiency
 - Hash table indexing enables $O(1)$ search time complexity



Methodologies (Cont'd)

Post-OCR Correction

Neural Machine Translation (NMT) Approaches

- BiLSTM Network:
 - Captures long-range dependencies in both forward and backward directions for leveraging contextual relationships for effective error detection and correction
- Transformer Architecture:
 - Uses self-attention mechanism for global context-awareness and processes input sequences simultaneously, overcoming limitations of recurrent networks

Methodologies (Cont'd)

Post-OCR Correction

Large Language Model (LLM) Approaches

- mT5 (Multilingual T5-based) transformer model
 - Text-to-Text Transfer Transformer (T5)
 - mT5-base trained on 101 languages and leveraged a unified text-to-text format
 - Fine-tuned for OCR correction, handling diverse errors
- mBART (Multilingual BART-based model) for denoising and text generation
 - mBART-50 variant fine-tuned for multiple languages
 - Primarily aimed at being fine-tuned on translation tasks
 - Pre-trained with a denoising objective, ideal for noisy OCR data



Experiments

Optical Model Training

- **Dataset:** myOCR dataset with iterations at 3,000, 6,000, and 9,000
- **Hidden States:** Trained models with 64, 128, and 256 hidden states
- **Feature Extraction:** VGG feature extractor used for image-to-feature conversion
- **Feature Sequence Modeling and Prediction:**
 - Models trained with and without BiLSTM for sequence modeling
 - CTC Decoding: Applied for sequence prediction
- **Results:**
 - 3,000 Iterations: Unable to decode feature maps
 - 6,000 & 9,000 Iterations: Successfully decoded with some OCR errors

Experiments

Myanmar Language Structure:

Hierarchical composition

words

syllables

grapheme-clusters

characters

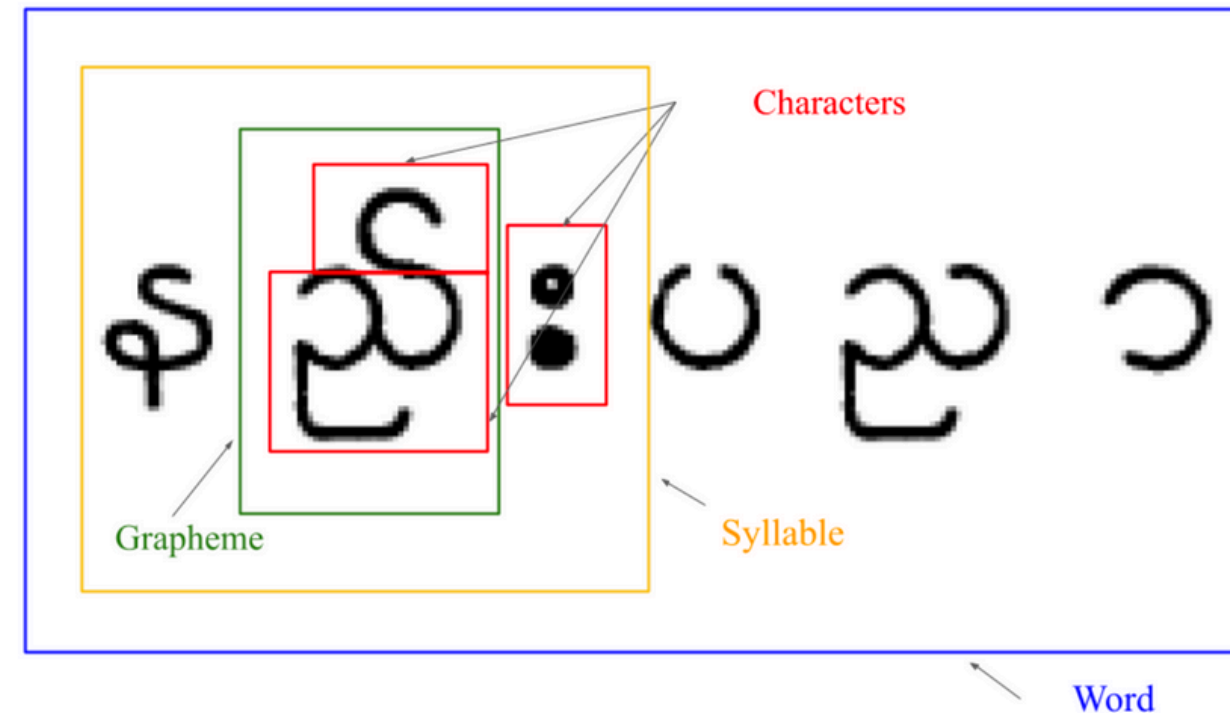



Fig. 2: Grapheme Composition of Myanmar word နည်းပညာ (“Technology” in English) in which Characters in Red, Grapheme-cluster in Green, Syllable in Yellow, and Word in Blue boxes



Experiments (Cont'd)

OCR Errors

- Conducted at the character level using SCLITE (NIST Scoring Toolkit)
- Aligns OCR hypothesis text with ground-truth to assess sequence- and word-level accuracy
- Confusion pairs identified for specific character-level errors

Experiments (Cont'd)


Confusion Pairs

TABLE III: SAMPLES FOR EACH ERROR TYPE RANDOMLY EXTRACTED FROM THE OPTICAL MODEL RESULTS

Error Type	Error and Ground-truth
Selection	<div>လ ၵ ဝံ ညီ ငွ ၵ ဝံ ရ ဝေ ဝး</div> <div>လ ၵ ဝံ ခ ငွ ၵ ဝံ ရ ဝေ ဝး</div>
Insertion	<div>က ငွ ည ဝံ န်</div> <div>က ငွ ည ဝံ ဝံ</div>
Deletion	<div>အ က ဝံ ဝံ ဝး မ ပ ဝေ ဝး</div> <div>အ က ဝံ ဝံ ဝး မ ပ ဝေ ဝး</div>

TABLE IV: THE TOP 10 VISUAL ERROR CONFUSION PAIRS FROM OCR MODELS (CORRECT → OCR ERROR)

Frequency	Confusion Pair
361	၁ → ၁
236	၀ → ၀
152	၀ → ၀
136	၁ → ၁
112	၀ → ၀
107	၁ → ၁
104	၁ → ၁
94	၀ → ၁
85	၁ → ၁
81	၁ → ၁



Experiments (Cont'd)

Post-OCR Experiments

Statistical Approaches


N-gram and SymSpell experiments:

- N-gram: Tested with N values of 3, 4, and 5.

Best result was obtained with character trigram (N=3).

- SymSpell: Tested with edit distances of 2, 3, and 4.

Optimal result found with edit distance 4.



Experiments (Cont'd)

Post-OCR Experiments

NMT Approaches

BiLSTM-based S2S:

- Max sequence length: 100 tokens.
- Hidden layers: 2, hidden size: 256.
- Batch size: 64, dropout: 0.5, early stopping after 10 validations.

Transformer-based S2S:

- Max sequence length: 100 tokens.
- Encoder/decoder: 4 layers, vector size 128, hidden size 128.
- Feed-forward dimension: 512, attention heads: 4.
- Batch size: 8, dropout: 0.1, learning rate 0.1.



Experiments (Cont'd)

Post-OCR Experiments

LLM Approach:

- Used mT5-base and mBART-50 for correction tasks.
- Sequence length: 96 tokens.
- Training:
 - mT5-base: 10 epochs, batch size 4.
 - mBART-50: 5 epochs, batch size 1.
- Mixed precision training (FP16) was used to manage memory limitations.



Experiments (Cont'd)

System Configuration

Optical Models Training System:

- AMD Ryzen 9 3900X
(12 cores, 24 threads)
- NVIDIA Quadro RTX 4000
(8 GB VRAM)
- 32 GB RAM

Post-OCR Correction Experiments System:

- Intel Core i9-14900KF (64-bit, 5701 MHz)
- 64 GB RAM
- NVIDIA RTX A6000 (48 GB GDDR6 VRAM)
- 300W power consumption

Experiments (Cont'd)

Evaluation Metrics

Word Error Rate (WER)

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

The metric for evaluating text recognition systems, such as OCR or speech recognition, based on the percentage of words that require insertion, deletion, or substitution to align the predicted text with the ground truth.

Character N-gram F-score (CHRF++)

$$\text{CHRF}\beta = (1 + \beta^2) \frac{\text{CHRP} \cdot \text{CHRR}}{\beta^2 \cdot \text{CHRP} + \text{CHRR}}$$

This metric calculates the F-score for predicted text sequences at the character N-gram level, focusing on precision and recall.

Experiments (Cont'd)

TABLE V: WORD ERROR RATE (WER) SCORES FOR EACH MODEL. THE LOWER THE WER, THE BETTER THE MODEL IS.

	Iteration	3,000			6,000			9,000		
	Hidden State	64	128	256	64	128	256	64	128	256
None	Base	N/A	N/A	N/A	30.20	23.92	35.47	23.46	18.94	19.15
	+N-gram	N/A	N/A	N/A	14.65	10.03	17.25	9.42	7.16	7.40
	+SymSpell	N/A	N/A	N/A	13.98	9.81	15.95	9.49	7.53	7.65
	+BiLSTM-S2S	N/A	N/A	N/A	15.07	10.83	19.60	10.20	7.54	7.54
	+Transformer-S2S	N/A	N/A	N/A	9.89	6.37	12.95	6.12	3.57	3.24
	+mT5-base	N/A	N/A	N/A	32.19	30.98	33.60	30.74	30.09	30.57
	+mBART-50	N/A	N/A	N/A	11.24	11.01	11.69	10.94	10.57	10.63
+BiLSTM	Base	15.87	17.32	12.30	10.74	11.05	9.69	10.04	10.11	9.18
	+N-gram	5.25	6.43	2.98	2.33	2.58	1.79	1.98	1.99	1.59
	+SymSpell	5.26	6.06	3.17	2.40	2.54	1.83	2.04	1.99	1.53
	+BiLSTM-S2S	7.05	6.56	5.31	4.58	4.27	3.83	4.10	4.05	3.71
	+Transformer-S2S	3.65	3.35	1.69	1.32	1.36	0.83	1.05	1.00	0.66
	+mT5-base	30.41	30.68	29.73	29.56	29.75	29.51	29.69	29.61	29.54
	+mBART-50	10.73	10.87	10.64	10.54	10.61	10.62	10.60	10.61	10.64

Experiments (Cont'd)

TABLE VI: CHARACTER N-GRAM F-SCORES (chrF^{++}) FOR EACH MODEL. THE HIGHER THE chrF^{++} VALUE, THE BETTER THE MODEL IS.

	Iteration	3,000			6,000			9,000		
	Hidden State	64	128	256	64	128	256	64	128	256
None	Base	N/A	N/A	N/A	80.54	86.17	75.09	86.76	90.32	89.86
	+N-gram	N/A	N/A	N/A	88.03	92.19	84.98	92.39	94.66	94.24
	+SymSpell	N/A	N/A	N/A	88.84	92.22	86.36	92.26	94.28	93.99
	+BiLSTM-S2S	N/A	N/A	N/A	83.56	87.84	79.30	88.69	91.34	90.98
	+Transformer-S2S	N/A	N/A	N/A	90.84	94.15	87.93	94.45	96.89	96.85
	+mT5-base	N/A	N/A	N/A	70.46	71.48	69.48	71.33	71.89	71.53
	+mBART-50	N/A	N/A	N/A	89.23	89.50	88.89	89.52	89.81	89.76
+BiLSTM	Base	91.81	91.25	95.26	96.55	96.14	97.53	97.21	97.05	97.90
	+N-gram	95.31	94.32	97.36	97.91	97.63	98.40	98.24	98.19	98.54
	+SymSpell	95.20	94.82	97.32	97.85	97.71	98.37	98.17	98.19	98.31
	+BiLSTM-S2S	91.14	91.49	93.18	93.98	93.96	94.52	94.26	94.25	94.58
	+Transformer-S2S	96.10	96.56	98.25	98.81	98.58	99.07	98.94	99.00	99.31
	+mT5-base	71.43	71.44	72.19	72.15	72.05	72.19	71.97	72.04	72.11
	+mBART-50	89.65	89.51	89.79	89.82	89.77	89.77	89.80	89.82	89.80



Experiments (Cont'd)

Experimental Results

Statistical - N-gram and SymSpell:

- N-gram and SymSpell show significant reductions in Word Error Rate (WER).
- Feature sequence modeling before correction reduces WER nearly fivefold, with statistical correction where N-gram outperforms SymSpell slightly.

Neural Machine Translation (NMT):

- BiLSTM-S2S and Transformer-S2S reduce WER, with Transformer-S2S yielding the best results by capturing context better.

Large Language Models (LLMs):

- mT5 and mBART are less effective, with mT5 struggling to reduce WER and mBART performing better at lower iterations (3,000) but less so at higher ones (6,000, 9,000).



Conclusion

- **Model Evaluation:** We tested OCR models with varying iterations and hidden states, applying statistical, NMT, and LLM-based Post-OCR corrections. Error analysis showed CTC decoding prioritized visual order over writing order.
- **Key Findings:** Transformer-S2S models showed superior performance, especially in handling minimal OCR errors across different iterations.



Future Work

- **Real-World Testing:** Evaluate on manually annotated datasets with more real-world variability.
- **Hyperparameter Optimization:** Tuning maximum sequence length and epoch for enhanced performance.
- **Opensourcing:** Release dataset and OCR models with Post-OCR configurations to support Myanmar language research.



Thank You.

Q & A

Reference

- Doermann (1998): Indexed and retrieved document images, Computer Vision and Image Understanding, Vol. 70(3), 287-298.
- El-Sawy et al. (2017): Recognized Arabic handwritten characters using CNNs, WSEAS Transactions on Computer Research, Vol. 5.
- Alwzwazy et al. (2016): Handwritten digit recognition via CNNs, Int. J. Innov. Res. Comp. Commun. Eng., Vol. 4(2), 1101-1106.
- Wang et al. (2012): End-to-end text recognition using CNNs, Proc. ICPR.
- Schuster & Paliwal (1997): Bidirectional RNNs, IEEE Trans. Signal Processing, Vol. 45(11), 2673-2681.
- Dutta et al. (2018): Hybrid CNN-RNN for handwriting recognition, Proc. ICFHR.
- Graves et al. (2009): Novel system for unconstrained handwriting recognition, IEEE TPAMI, Vol. 31, 855-868.

Reference (Cont'd)

- Shi et al. (2017): Trainable NN for image-based sequence recognition, IEEE TPAMI, Vol. 39, 2298-2304.
- Graves et al. (2006): CTC for unsegmented sequence data, Proc. ICML, 369-376.
- Stefanović et al. (2019): N-grams text similarity with self-organizing maps, Applied Sciences, Vol. 9, Art. no. 1870.
- Garbe (2012): Spelling correction with SymSpell algorithm, Towards Data Science.
- Kayabas et al. (2021): OCR error correction using BiLSTM, Proc. ICECET, 1-5.
- Yasin et al. (2023): Transformer for post-OCR error correction, Proc. ICDAR Workshops, 80-93.
- Madarász et al. (2024): OCR cleaning of scientific texts using LLMs, NSLP, Vol. 14770.
- Soper et al. (2021): BART for OCR newspaper text post-correction, Proc. W-NUT.

Reference (Cont'd)

- Hlaing et al. (2022): POS-tag features in NMT for low-resource languages, Heliyon.
- Simonyan & Zisserman (2015): Deep CNNs for image recognition, Proc. ICLR, Vol. 4.
- Hochreiter & Schmidhuber (1997): LSTM, Neural Computation, Vol. 9(8), 1735-1780.
- Vaswani et al. (2017): Transformer model, NIPS 2017.
- Xue et al. (2021): mT5 multilingual text-to-text transformer, Proc. NAACL, 483-498.
- Liu et al. (2020): Multilingual pre-training for NMT, TACL, Vol. 8, 726-742.
- Mon et al. (2021): SymSpell4Burmese for Burmese spelling correction, Proc. iSAI-NLP.
- NIST: SCKT - Scoring Toolkit, GitHub.
- Hosken & Tuntunlwin (2004): Unicode representation of Myanmar, Semantic Scholar.