myNLP: Natural Language Processing Library for Myanmar Language

Anonymous ACL submission

Abstract

myNLP is a free, open-source natural language processing (NLP) library focused on the Myanmar language. The library is implemented in Python programming language and benchmarked on the available Myanmar corpora. In this paper, we provide outlines and comparisons of different approaches for each of the language processing functionalities as well as the datasets and pre-trained models. The library is constructed in a hierarchical structure including language processing functions and models for different NLP tasks. It will be publicly released and available on GitHub, with some larger models hosted on Hugging Face.

1 Introduction

In recent years, there have been many advancements in the Natural Language Processing (NLP) field with the advent of Multilingual Language Models (MLLMs) and Large Language Models (LLMs). Although there are many NLP tools and libraries, most of them are designed for languages with many resources. With the limited amount of data for the experiments, low-resource languages were left behind.

Myanmar language, which is a low-resource language is the official language of the Republic of the Union of Myanmar (Constitution of Myanmar (2008)) and is spoken by two-thirds of the population (SIL International (2024)). Despite its significance, there is a lack of a comprehensive NLP toolkit for the Myanmar language. Therefore, in this paper, we aim to introduce an NLP library for the Myanmar language which will fill the gap for many language processing tasks.

In myNLP, we included important preprocessors for linguistic analysis and pre-trainings, such as tokenization, part-of-speech tagging, and name-entity recognition. In the following sections, we describe the design and implementation of our library. We also evaluate the performance of our library on various tasks and compare it with the earlier studies.

2 Functionalities

We used both rule-based and data-driven approaches for our library. Data-driven models were trained on the open-source corpus as described in Table 1. The experiment details will be discussed in the next sections.

2.1 Tokenization

myNLP supports text tokenization for different units including grapheme clusters, syllables, words, and sentences.

2.1.1 Grapheme Clusters Tokenization

ICU (International Components for Unicode) (IBM Corporation et al. (1999)) is a library that provides robust and efficient Unicode support. ICU grapheme segmentation involves dividing text into grapheme clusters, which are sequences of one or more Unicode code points that represent a single user-perceived character. Grapheme clusters are the atomic units of matching in Unicode. We used PyICU[1] which is the python extension implemented in C++ that wraps the C/C++ ICU library.

2.1.2 Syllable Tokenization

Myanmar language characters (consonants, vowels, and diacritic marks) form Myanmar language syllables. Since the Myanmar language is monosyllabic, i.e., the syllable is the unbreakable unit in the Myanmar language, syllable segmentation is important for Myanmar language text processing. Thu et al. (2021) proposed the syllable-breaking regular expression patterns for Nine major ethnic languages of Myanmar in Perl. We implemented the authors' proposal for the Myanmar language syllable tokenization structure in Python to include it in our library.

[1] https://pypi.org/project/PyICU/

2.1.3 Word Tokenization

In the Myanmar language, a word consists of one or more syllables (Green (2005)). myNLP supports three-word tokenization approaches based on n-gram dictionaries with myWord Thu (2021), machine learning with Conditional Random Fields (CRF) (Fukuda et al. (2007)), and deep learning with Bidirectional Long Short-Term Memory (Bi-LSTM) (Ma et al. (2018)). Ma et al. (2018) compares various neural network architectures for Chinese word segmentation and finds that a Bidirectional Long Short-Term Memory (Bi-LSTM) model has better accuracy results compared to other models. From the orthographical point of view, since the Myanmar language and Chinese have the common intrinsic problem of defining the word boundary (Ding et al. (2016)), we used deep sequence labeling as in Ma et al. (2018).

2.1.4 Phrase Tokenization

myNLP also supports phrase tokenization for phrase-based NLP tasks. Our phrase tokenizer is built on the myWord tool using the unsupervised approach with Normalized Pointwise Mutual Information (NPMI) proposed by Bouma (2009).

2.1.5 Sentence Tokenization

Sentence Tokenization is useful for various kinds of NLP applications such as machine translation, automatic speech recognition, and information extraction. Aung et al. (2023) proposed mySentence, the first sentence segmentation corpus with RDR, CRF, and Hidden Markov Models (HMM) tagging methods compared with Neural Machine Translation (NMT) approaches. It was found that neural sequence tagging experiments made by Thu et al. (2023a) outperform the traditional tagging and NMT methods. Our library developed a word-level neural sequence labeling model using Bi-LSTM architecture and it is trained and evaluated on mySentence corpus.

2.2 Tagging

myNLP supports Part-Of-Speech (POS) Tagging and Name-Entity Recognition (NER), which are important language processing methods for linguistic analysis and information retrieval.

2.2.1 POS Tagging

POS Tag represents the syntactic category of the word. We used myPOS version 3 with 16 POS Tags defined by Hlaing et al. (2022b). The authors proposed that the Ripple-Down Rules (RDR) POS Tagger has better results than neural network approaches. For myNLP, we implemented three different approaches - CRF, RDR, and Bi-LSTM by training and evaluating on the same corpus.

2.2.2 Name-Entity Recognition

Name-Entity Recognition in myNLP can provide name entity information for the users with 9 Name-Entity tags with BIOES tagging scheme. We trained and implemented CRF and Bi-LSTM models on our developing myNER corpus version 1.0. We plan to release it together with the myNLP framework.

2.3 Transliteration

Myanmar language is a tonal language with four tones: low, high, creaky, and stopped (checked). Varied tones impart distinct meanings to syllables with identical phonemic structures. Therefore phonetic information are important for Myanmar language linguistics studies. For the further phonological studies, myNLP contributes the Transliteration and Grapheme-to-Phoneme conversion module.

myNLP supports two types of mapping-based transliteration methods (Sawada (2021) and ALA-LC (2011)) and two Bi-LSTM models for Grapheme-to-Phoneme (G2P) conversion and Grapheme-to-IPA (G2IPA) conversion. Bi-LSTM conversion models are trained and evaluated on myG2P word-level dictionary version 2.0 developed by Htun et al. (2021).

2.4 Dependency Parsing

Universal Dependency (UD) parsing is used in various kinds of NLP tasks especially for language understanding and generation by providing a structured representation of the relationships between words. myNLP provides both graph-based and transition-based dependency parsing models.

2.5 Spelling Checking

Spelling error detection and correction are crucial components of text normalization for training contextual models. myNLP offers word-level spelling correction utilizing the SymSpell algorithm developed by Garbe (2012). Building upon previous research by native experts, we have adopted the most promising approach, specifically the Damerau-Levenshtein method, along with unigram and bi-

gram dictionaries manually curated and proposed by novice experts (Mon et al. (2021)).

2.6 String Similarity

Wai et al. (2020) presented the development of string similarity measures based on phoneme similarity. Based on their study, myNLP provides string similarity scores for all kinds of mappings - phonetic, sound, and vowel position mappings with various kinds of edit distance including Levenshtein distance, Damerau-Levenshtein distance, Hamming distance, Jaro-Winkler distance, cosine similarity, and Jaccard distance.

2.7 Paraphrase Classification

We implemented the Random Forest (RF) classifier, Siamese Convolutional Neural Network (CNN), and Bi-LSTM conducted in the research of Htay et al. (2022). We used RF as our paraphrase classification model because it outperformed the performance of the Siamese models.

2.8 Text Classification

We included the hate speech detection function as a part of our text classification module. We implemented and tested traditional Machine Learning (ML) algorithms such as Support Vector Machine (SVM), Multinomial Naive bayes (NB), and Random Forest (RF) (Marshan et al. (2023)), as well as fasttext classification model (Joulin et al. (2016)).

2.9 Language Classification and Embeddings

Since there is no language classification model released between the Myanmar language and other eight ethnic languages (Beik, Dawei, Mon, Pao, Po-Kayin, Rakhine, Sgaw-Kayin, and Shan), myNLP provides a language classification module to improve local low-resource language identification.

- **Classification**: We included a language classification function using character and syllable n-grams with Naive Bayes (Vatanen et al. (2010)), allowing users to choose from n-values of 3, 4, and 5. Another method for language classification is using character-syllable frequencies for each ethnic language. We also trained neural network model (Accuracy: 99.5%) and fasttext classification models (Joulin et al. (2016)) (Accuracy: 99.7%) to classify the language of the input string.

- **Syllable Embeddings**: myNLP will also release the syllable-level embeddings for the nine ethnic languages including the Myanmar language to be used in various NLP tasks such as language classification, semantic similarity, and machine translation. The ethnic language corpus is segmented into syllables using the sylbreak4all tool (Thu et al. (2021) and trained fasttext (Bojanowski et al. (2017)) and word2vec (Mikolov et al. (2013)) on the monolingual corpora using gensim[2].

2.10 Machine Translation

In order to conduct research and development (R&D) on machine translation for local languages, members of the myNLP team contribute to the development of parallel corpora between the Myanmar language and other major ethnic languages such as Kachin, Kayar, Pa'O, Rawang, Sgaw Kayin, and Shan, as well as parallel corpora between Burmese spoken dialects (Beik, Dawei, and Rakhine) (Kyaw et al. (2020), Thu et al. (2019);Oo et al. (2023)). Moreover, we also plan to release English-Myanmar parallel corpus for medical domain together with this myNLP library.

The myNLP framework supports both statistical machine translation (SMT) and neural machine translation (NMT). SMT utilizes IBM Models 1 and 2 (Koehn (2010)), tailored to accommodate translations for local spoken dialects and one of Myanmar's Braille systems known as Mu-thit Moe et al. (2021). On the other hand, myNLP's NMT implementation leverages OpenNMT-tf (Klein et al. (2018); Klein et al. (2020)) and CTranslate2[3] libraries to facilitate efficient inference with Transformer models.

2.11 Utilities

We also included a utilities module for language processing tasks such as corpus cleaning, normalization, language modeling, stopword removal (Thu and Supnithi (2023)), sorting, etc.

3 Datasets

myNLP also provides datasets for the researchers to be able to use in their further Myanmar linguistics research. Table 1 shows the dataset available in myNLP ecosystem. The first row in each dataset is the train set and the second row is the test set. Some datasets are not sentence-level corpus but word-level dictionaries and described as N/A meaning that these were not developed with sentences.

[2] https://pypi.org/project/gensim/
[3] https://github.com/OpenNMT/CTranslate2

3.1 Word Segmentation corpus

Word segmentation corpus is manually checked and segmented corpus by the native Myanmar language speakers. Since there is no publically available word segmentation corpus, we developed our word segmented corpus using data collected from various domains including social media websites and news websites. For the further tokenization experiments, we tagged our corpus based on the proposal of Pa et al. (2015) in both character and syllable levels.

3.2 mySentence

mySentence is annotated using word only part of myPOS version 3.0 and additional sentences from the internet resources (Aung et al. (2023)). The authors annotated word sequences in the corpus into a tagged sequence of words. Each token within the sentence was assigned one of the four tags: B (Begin), O (Other), N (Next), or E (End).

3.3 myPOS version 3.0

According to Htike et al. (2017), myPOS version 1.0 contains 11,000 sentences collected from Wikipedia[4].

myNLP supports both myPOS and Univers POS (UPOS) Tag, which are used in the UD framework configurations. myPOS contains 15 tag sets - abb (Abbreviation), adj (Adjective), adv (Adverb), conj (Conjunction), fw (Foreign word), int (Interjection), n (Noun), num (Number), part (Particle), ppm (Post-positional Marker), pron (Pronoun), punc (Punctuation), sb (Symbol), tn (Text Number), and v (Verb). UPOS Tags were defined by Petrov et al. (2012) as NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners & articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuations), and X (for other categories).

3.4 myNER version 1.0

Since there is no open-source NER corpus for Myanmar language, we developed our own NER corpus using 9 tag sets - "LOC: Location", "EVENT: Event", "DATE: Date", "PER: Person", "NUM: Number", "PRODUCT: Product", "TIME: Time", "ORG: Organization" and "O: Outside" using BIOES tagging scheme. It is annotated based on myPOS ver 1.0 corpus and additional sentences from Wikipedia. Similar to mySentence and myPOS, myNER also tagged in word-level.

3.5 myG2P version 2.0

myG2P version 2.0 is extened version of myG2P version 1.0 and 1.1 dictionaries. It includes IPA column which was not in earlier versions. The dictionary dataset consists of 2,353 unique syllables, 1,928 unique IPA symbols and a total of 24,803 G2IPA pairs. The syllables were modified based on Unicode (version 13.0). myG2P was used in VoiceTra[5] (Multilingual Speech Translation Application) project of NICT, Japan (during 2014-2015).

3.6 myParaphrase version 1.0

Htay et al. (2022) conducted a semantic similarity classification for the Myanmar language and developed the first paraphrase classification corpus. Paraphrase classification corpus contains more than 41K pair sentences with paraphrase or not binary labels. myParaphrase corpus contains 15,640 paraphrase sentences and 24,821 non-paraphrase sentences. Therefore, it can be useful to build paraphrase classification as well as generation systems.

3.7 myRoman version 1.0

myRoman (Zaw et al. (2020)) which stands for Myanmar Romanization is a collection of romanized names in Myanmar country. Names were collected not only from the domain of Bamar ethnic, but also from Kachin, Kayah, Kayin, Chin, Mon, and Rakhine ethnics. The authors included every possible syllable for the names that (e.g. Hain, Haine, and Hein) share the same spelling Myanmar language. The corpus is segmented into syllables as in myG2P and contains 1,489 unique syllables in the corpus.

3.8 myPoetry version 1.0

myPoetry is a collection of Myanmar poems for creative computational poem generation ((Thu et al., 2023b)). The corpus is composed of 1,873 poems from 83 books written by 393 poets. The poems are in different styles from classical to modern and also contain translations. Chin and Rakhine (other ethnic languages of Myanmar) poems were also included in this version. The authors also released a statistical and finetuned GPT2 language models for poetry generation.

[4] https://my.wikipedia.org/wiki

[5] https://voicetra.nict.go.jp/en/index.html

Corpus	Tokens	Sentences
Word Segmentation	5,075,674 266,673	202,448 10,656
mySentence	896,025 96,632	50,081 5,512
myPOS ver 3.0	524,408 12,825	42,196 1,000
myNER ver 1.0	212,563 23,745	13,762 1,530
myG2P ver 2.0	22,324 2,481	N/A N/A
myParaphrase ver 1.0	591,452 10,706	40,461 1,000
myRoman ver 1.0	50,111 5,000	N/A N/A
myPoetry ver 1.0	100,676	46,933
myUDTree ver 1.0	564,505	43,196
English ↔ Myanmar	264,636 17,917	13,133 1,459
myHateSpeech	413,800	20,280

Table 1: Datasets available in myNLP.
First row in each dataset is for the training dataset and the second row for the test dataset. N/A indicates that the datasets are token-level datasets.

3.9 myUDTree version 1.0

The myUDTree corpus (Hlaing et al. (2022a)), an extension of prior Myanmar UD Corpus (Aye et al. (2018)), comprises 43,196 sentences, enhancing the Myanmar UD corpus by incorporating 11K sentences of dependency tree data.

There are 14 Universal Part-of-Speech tags and 14 dependency relations applied such as root, acl (clausal modifier of noun), amod (adjectival modifier), advmod (adverbial modifier), case (case marking), mark (marker), compound (compound), obl (oblique nominal), obj (object), and punct (punctuation). The CoNLL-U format serves as the chosen dependency-tree format for myUDTree corpus.

3.10 English - Myanmar Parallel corpus

San et al. (2024) developed an English-Myanmar parallel corpus focused on the medical domain for low-resource NLP machine translation research. The corpus consists of more than 14 thousand sentences and is segmented into syllables.

3.11 myHateSpeech

myHateSpeech is a binary classification dataset with word-segmented sentences to classify hate speech or not. The data were collected from social media websites such as Facebook[6] and labeled manually by the native speakers.

4 Methodologies

In this section, we discuss the overview of the algorithms used in myNLP library. Deep learning framework Tensorflow (Abadi et al. (2015)) is used to build and train deep learning models. The hyperparameters for each are described in Table 2.

4.1 N-grams

N-grams are fundamental in capturing the sequential nature of language and have been integrated into various language models, classification algorithms, and information retrieval systems. We applied the concept of n-grams in myNLP for word segmentation and language classification tasks.

myWord is a tool developed by Thu (2021) and released open-source along with the unigram and bigram dictionaries for syllable, word, and phrase segmentation for the Myanmar language. For word segmentation, the author used a corpus with more than 0.5M sentences and 12M words to generate unigram and bigram dictionaries. The characters from the unsegmented sentence are scored based on the dictionaries and decoded using the Viterbi algorithm (Viterbi (1967)) to get the most possible combination. During scoring, the probability in the bigram is picked if the combination is in the bigram dictionary. If not, the probability from the unigram dictionary will be taken.

For language classification, in other words, language identification, we leveraged the multinomial Naive Bayes classifier with character and syllable n-grams as proposed in ((Vatanen et al. (2010))).

4.2 RDR Tagger

Ripple Down Rules (RDR) is an approach to building knowledge-based systems incrementally while they are already in use. It involves the creation of transformation rules in the form of Single Classification Ripple Down Rules (SCRDR) based on the concept of incremental and case-based knowledge acquisition (Nguyen et al. (2014)).

[6] https://www.facebook.com/

Model	Hyperparameters	
CRF	Optimization:	L-BFGS
	L1 regularization:	1.0
	L2 regularization:	1e-3
	Maximum Iteration:	100
	Transition Features:	Enabled
Bi-LSTM	Embedding Dimension:	100
	Learning rate:	1e-3
	Batch Size:	64
	Hidden Units	50
	Optimization:	Adam
	Activation:	Softmax
	Epoch:	30
Fasttext	Embedding Dimension:	100
	Minimum Count:	1
	Word N-grams:	6
	Character N-grams	3-6
	Learning Rate	1e-1
	Context Window:	5
	Activation:	Softmax
Siamese	Embedding Dimension:	100
	Learning rate:	1e-3
	Batch Size:	512
	CNN Filters	50
	Hidden Bi-LSTM Units	128
	Optimization:	Adam
	Epoch:	10
Fast DPNN	Embedding Size:	50
	Dropout Rate:	0.4
	Hidden Units:	100
	Learning Rate:	1e-3
	Batch Size:	64
	Optimization:	Adam
	Epoch:	30
S2S	Dropout Rate:	0.1
	Batch Size:	64
	Optimization:	Adam
	Parameter Sharing:	False
Biaffine	Embedding Size:	100
	Dropout Rate:	0.33
	Bi-LSTM Units:	256
	arc MLT size:	500
	rel MLT size:	100
	Optimization:	Adam
	Learning Rate:	2e-3
	Batch Size:	64
	Epoch:	30

Table 2: Hyperparameters of myNLP models

4.3 CRF

CRF are a type of probabilistic graphical model that can be used for various tasks such as sequence tagging, including Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and token classification tasks like Grapheme-to-Phoneme (G2P) and Grapheme-to-IPA (G2IPA) conversion. CRF can also be used to learn several probabilistic parameters from the training data to predict word boundaries. We can also detect sentence boundaries using CRF by tagging words as sentence boundary tags. For Myanmar language word segmentation, we experimented using the feature sets defined by Pa et al. (2015). We used pyCRFuite[7] software to train the CRF models on our datasets.

4.4 Bi-LSTM

Bi-LSTM is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backward direction. Bi-LSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm. They are commonly used for tasks such as token classification, NER, and POS tagging. The Bi-LSTM model is based on the LSTM unit and can effectively process long-sequence data and long-term model dependencies. It is well suited for tasks that require understanding the context from both preceding and following words, making it a suitable architecture for various natural language processing tasks.

4.5 Siamese Neural Networks

The Siamese CNN and Siamese RNN are both neural network structures commonly used in paraphrase detection also known as semantic similarity detection tasks. While the Siamese CNN focuses on gauging the semantic likeness of two sentences, the Siamese RNN utilizes recurrent neural network (RNN) layers for sequential data processing, serving a similar purpose in paraphrase detection (Ranasinghe et al. (2019)).

4.6 Dependency Parsing models

We trained the greedy, transition-based neural network parser called Fast and Accurate Dependency Parser using Neural Networks (Fast Accurate DPNN) (Chen and Manning (2014)), biaffine (Dozat and Manning (2016)) and jPTDP (Nguyen and Verspoor (2018)) models. While training, the

[7] https://pypi.org/project/python-CRFuite/

myUDTree corpus is split into train(70%), validation(10%) and test(20%) datasets.

The biaffine model is a graph-based dependency parser that uses neural attention and biaffine classifiers to predict arcs and labels, achieving state-of-the-art performance on standard treebanks. jPTDP is a neural network based joint POS tagging and dependency parsing model. Python 2.7 and DyNet[8] software were required to train the jPTDP model. It is also used in spaCy[9], a popular language processing library.

5 Results and Discussion

Accuracy (Acc), Precision (P), Recall (R) and F1-Score (F1) were used to evaluate the performance of each classification and sequence labeling model. It is calculated as the ratio of the number of correct predictions to the total number of predictions. Precision measures the accuracy of positive predictions. Recall measures the proportion of actual positives that were correctly predicted. And F1-Score is the harmonic mean of precision and recall. The higher the scores, the better our model. Table 3, 4, 5 and 6 compares the performance of the models using P, R, Acc and F1 scores.

To evaluate the dependency parsing models, we use UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score). UAS measures the proportion of words for which the parser correctly assigns a head, regardless of the specific dependency label. On the other hand, LAS considers both the correct assignment of a head and the accurate labeling of the dependency. Table 7 shows the comparison of our dependency parsing models with jPTDP conducted by Hlaing et al. (2022a).

- **Tokenization**: For word and sentence tokenization, we implemented CRF and Bi-LSTM models. Table 3 shows the performance of myWord presented by Thu (2021), and Accuracy and F1-score of CRF and Bi-LSTM models to detect word boundary tag ("|") with different tag set configurations. For the word segmentation, myWord has accuracy of 0.98 and F1-score of 0.88 while CRF model gained 0.96 Acc and 0.97 F1 with 2-Tag syllable based configuration and Bi-LSTM gained both Acc and F1 0.98 with 4-tag syllable.

- **Tagging**: For the tagging experiments, RDR Tagger is the best model for POS Tags (Hlaing et al. (2022b)). For NER, although the accuracy is good for Bi-LSTM, other scores still show that training data of NER is highly imbalanced.

- **Token Classification**: Thu et al. (2016) and Htun et al. (2021) compare the different algorithms. According to our experiments in Table 4, performances of both G2P and G2IPA have no significant difference. CRF models give the best results compared with RDR Tagger and Bi-LSTM models.

- **Text Classification**: Traditional ML (SVM, NB, and RF) and fasttext models were trained on the myHateSpeech data for binary classification with both syllable and word units. TF-IDF vectorization method with n-gram range of 1 to 3 was used to train ML models and the hyperparameters for fasttext model are as shown in the Table 2. We found out that syllable segmentation improved text classification and fasttext gained the best performance of 96% accuracy with syllable tokens (Table 5) on the test data which is 20% of the corpus.

- **Machine Translation**: Sequence to Sequence (S2S) model was trained on our Medical English-Myanmar parallel corpus and gained the BLEU score 31 with the hyperparameters described in Table 2.

- **Paraphrase Classification**: We trained Siamese neural networks of CNN and Bi-LSTM with the hyperparameters described in Table 2. Similar to the experiments conducted by Htay et al. (2022), both models gained good results on the closed test set but bad results on the open dataset. RF Classifier is also trained and we confirmed that the authors' results gained the same value on opened test data. The RF with the features from the Harry tool (Rieck and Wressnegger (2016)) outperformed the results of CNN and Bi-LSTM as the authors proposed. Table 6 shows the results of our paraphrase classification models.

- **Dependency Parsing**: Although the biaffine dependency parser we trained gained better results than the best resulted model - jPTDP by Hlaing et al. (2022a), unlike jPTDP, our model still depends on the performance of POS tagging. jPTDP does not need to have good

[8] https://github.com/clab/dynet
[9] https://pypi.org/project/spacy-jptdp/

POS Tagger since itself is the joint model trained for both POS Tagging and dependency parsing as well as for other columns. Therefore, tokenized raw text is enough to use the jPTDP. Since our RDR Tagger have the significant good result with myPOS version 3.0 corpus, we decided to use the biaffine model as myNLP's dependency parser.

6 Future works

This paper introduces the myNLP library, explaining its features, dataset development process, and showcasing some of the current research and development results achieved over a decade of effort. By 2024, we aim to implement and release an open-source library of fundamental NLP tasks such as word segmentation, sentence breaking, POS tagging, NER tagging, G2P conversion, UDTree parsing, Romanization, spelling checking, hate speech classification, GPT-2-based language model for poetry, and machine translation for Myanmar languages. Additionally, ongoing work includes speech corpus development, as well as ASR (Automatic Speech Recognition) and TTS (Text-to-Speech) modeling for the forthcoming version. In the near future, our plans involve extending this library to support more ethnic languages of Myanmar.

Acknowledgements

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

ALA-LC. 2011. ALA-LC Romanization Tables.

Thura Aung, Ye Kyaw Thu, and Zar Zar Hlaing. 2023. mysentence: Sentence segmentation for myanmar language using neural machine translation approach. *Journal of Intelligent Informatics and Smart Technology*, 9(October):e001. [Internet]. Available from: https://ph05.tci-thaijo.org/index.php/JIIST/article/view/87.

Hnin Thu Zar Aye, Win Pa Pa, and Ye Kyaw Thu. 2018. Unsupervised dependency corpus annotation for myanmar language. In *Proceedings of The 21st Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (Oriental COCOSDA 2018)*, Miyazaki, Japan.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gerlof J. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Constitution of Myanmar. 2008. Constitution of the republic of the union of myanmar. Chapter XV, Provision 450.

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita. 2016. Word segmentation for burmese (myanmar). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(4):1–10. Published on 16 May 2016.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734.

T. Fukuda, M. Izumi, and T. Miura. 2007. Domain dependent word segmentation based on conditional random fields. In *2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 268–271, Victoria, BC, Canada.

Wolf Garbe. 2012. Symspell.

Antony Dubach Green. 2005. Word, joot, and syllable structure in burmese. In Justin Watkins, editor, *Studies in Burmese Linguistics*. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University, Canberra ACT 0200, Australia.

Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul. 2022a. Graph-based dependency parser building for myanmar language. In *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul. 2022b. Improving neural machine translation with pos-tag features for low-resource language pairs. *Heliyon*, 8(8):e10375.

	2-Tag				3-Tag				4-Tag			
	character		syllable		character		syllable		character		syllable	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CRF	0.96	0.93	**0.96**	**0.97**	0.95	0.95	0.96	0.97	0.94	0.95	0.96	0.97
Bi-LSTM	0.98	0.97	0.98	0.97	0.97	0.97	0.96	0.98	0.95	0.96	**0.98**	**0.98**
myWord	Prec:	0.86	Re:	0.91	F1:	0.88	Acc:	0.98				

Table 3: Accuracy (Acc) and F1-Score (F1) of word boundary tag and performance of myWord

	RDR				CRF				Bi-LSTM			
	P	R	A	F1	P	R	A	F1	P	R	A	F1
Sentence	0.75	0.76	0.76	0.76	0.76	0.75	0.75	0.76	**0.96**	**0.96**	**0.95**	**0.96**
POS	**0.95**	**0.97**	**0.97**	**0.96**	0.95	0.91	0.96	0.93	0.94	0.93	0.95	0.94
NER	0.64	0.50	0.98	0.54	0.72	0.54	0.98	0.60	**0.70**	**0.55**	**0.99**	**0.59**
G2P	0.82	0.84	0.85	0.85	**0.98**	**0.98**	**0.99**	**0.98**	0.93	0.93	0.96	0.93
G2IPA	0.84	0.84	0.86	0.85	**0.97**	**0.99**	**0.99**	**0.98**	0.93	0.92	0.96	0.93

Table 4: Performace of Token Classification and Sequence labelling models in myNLP

	Word		Syllable	
	Acc	F1	Acc	F1
SVM	0.77	0.76	0.83	0.84
NB	0.75	0.73	0.81	0.81
RF	0.75	0.76	0.83	0.82
fasttext	**0.88**	**0.88**	**0.96**	**0.96**

Table 5: Performace of Text Classification models for Hate Speech Detection

Model	Close	Opened
Siamese CNN	0.99	0.47
Siamese Bi-LSTM	0.99	0.47
Random Forest	**0.99**	**0.85**

Table 6: Accuracy of Paraphrase Classification models on Closed/Opened Test data

Model	UAS	LAS
jPTDP	86.16%	82.77%
Fast DPNN	88.70%	85.00%
Biaffine	**94.84%**	**92.72%**

Table 7: UAS and LAS scores for UD parsing

Myint Myint Htay, Ye Kyaw Thu, Hnin Aye Thant, and Thepchai Supnithi. 2022. Deep siamese neural network vs random forest for myanmar language paraphrase classification. *Journal of Intelligent Informatics and Smart Technology*, 2(October 2nd Issue):25–1–25–9. (Submitted Feb 21, 2022; accepted July 17, 2022; published on 31 Oct 2022).

Khin War War Htike, Ye Kyaw Thu, Zuping Zhang, Win Pa Pa, Yoshinori Sagisaka, and Naoto Iwahashi. 2017. Comparison of six pos tagging methods on 10k sentences myanmar language (burmese) pos tagged corpus. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, Budapest, Hungary.

Honey Htun, Ni Htwe Aung, Shwe Sin Moe, Wint Theingi Zaw, Nyein Nyein Oo, Thepchai Supnithi, and Ye Kyaw Thu. 2021. Grapheme-to-ipa phoneme conversion for burmese (myg2p version 2.0). *Journal of Intelligent Informatics and Smart Technology*, 1(1).

IBM Corporation et al. 1999. International components for unicode. https://icu.unicode.org.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Guillaume Klein, François Hernandez, Vincent Nguyen,

and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, USA.

Nang Aeindray Kyaw, Ye Kyaw Thu, Hlaing Myat Nwe, Phyu Phyu Tar, Nandar Win Min, and Thepchai Supnithi. 2020. A study of three statistical machine translation methods for myanmar (burmese) and shan (tai long) language pair. In *Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2020)*, pages 218–223, Bangkok, Thailand.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art Chinese word segmentation with Bi-LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908, Brussels, Belgium. Association for Computational Linguistics.

Alaa Marshan, Farah Nasreen Mohamed Nizar, Athina Ioannou, and Konstantina Spanaki. 2023. Comparing machine learning and deep learning techniques for text analytics: Detecting the severity of hate comments online. *Information Systems Frontiers*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Zun Hlaing Moe, Thida San, Ei Thandar Phyu, Hlaing Myat Nwe, Hnin Aye Thant, Naw Naw, Htet Ne Oo, Thepchai Supnithi, and Ye Kyaw Thu. 2021. Myanmar text (burmese) and braille (mu thit) machine translation applying ibm model 1 and 2. *Journal of Intelligent Informatics and Smart Technology*, pages 18–26. Submitted February 8, 2021; accepted March 11, 2021; revised April 15, 2021; published online April 25, 2021.

Ei Phyu Phyu Mon, Ye Kyaw Thu, Than Than Yu, and Aye Wai Oo. 2021. Symspell4burmese: Symmetric delete spelling correction algorithm (symspell) for burmese spelling checking. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Dat Q. Nguyen, Dai Q. Nguyen, D. D. Pham, and S. B. Pham. 2014. Rdrpostagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20, Gothenburg, Sweden. Association for Computational Linguistics.

Dat Quoc Nguyen and Karin Verspoor. 2018. An improved neural network model for joint POS tagging and dependency parsing. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 81–91, Brussels, Belgium. Association for Computational Linguistics.

Thazin Myint Oo, Thitipong Tanprasert, Ye Kyaw Thu, and Thepchai Supnithi. 2023. Transfer and triangulation pivot translation approaches for burmese dialects. *IEEE Access*, 11:6150–6168.

Win Pa Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. 2015. Word boundary identification for myanmar text using conditional random fields. In *Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing (ICGEC 2015)*, pages 447–456, Yangon, Myanmar.

S. Petrov, D. Das, and R. McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2019. Semantic textual similarity with Siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, Varna, Bulgaria. INCOMA Ltd.

Konrad Rieck and Christian Wressnegger. 2016. Harry: A tool for measuring string similarity. *Journal of Machine Learning Research (JMLR)*, 17(9):1–5.

Mya Ei San, Sasiporn Usanavasin, Ye Kyaw Thu, and Manabu Okumura. 2024. A study for enhancing low-resource thai-myanmar-english neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Hideo Sawada. 2021. Sawada's roman transliteration system for burmese script. `http://www.aa.tufs.ac.jp/~sawadah/burroman2.pdf`. Language (ISO 639-3), Japanese.

SIL International. 2024. Ethnologue: Languages of the world. `https://www.ethnologue.com/language/mya/`.

Ye Kyaw Thu. 2021. myword: Syllable, word and phrase segmenter for burmese. `https://github.com/ye-kyaw-thu/myWord`.

Ye Kyaw Thu, Thura Aung, and Thepchai Supnithi. 2023a. Neural sequence labeling based sentence segmentation for myanmar language. In *The 12th Conference on Information Technology and its Applications (CITA 2023)*, volume 734 of *Lecture Notes in Networks and Systems*, pages 285–296, Danang, Vietnam. Springer, Cham.

Ye Kyaw Thu, Hlaing Myat Nwe, Hnin Aye Thant, Hay Man Htun, Htay Mon, May Myat Myat Khaing, Hsu Pan Oo, Pale Phyu, Nang Aeindray Kyaw, Thazin Myint Oo, Thazin Oo, Thet Thet Zin, and Thida Oo. 2021. sylbreak4all: Regular expressions for syllable breaking of nine major ethnic languages of myanmar. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Ye Kyaw Thu, Win Pa Pa, Yoshinori Sagisaka, and Naoto Iwahashi. 2016. Comparison of grapheme–to–phoneme conversion methods on a myanmar pronunciation dictionary. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), COLING 2016*, pages 11–22, Osaka, Japan.

Ye Kyaw Thu, Manar Hti Seng, Thazin Myint Oo, Dee Wom, Hpau Myang Thint Nu, Seng Mai, Thepchai Supnithi, and Khin Mar Soe. 2019. Statistical machine translation between kachin and rawang. In *Proceedings of the 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2019)*, pages 329–334, Chiang Mai, Thailand.

Ye Kyaw Thu and Thepchai Supnithi. 2023. Embedding meets frequency: Novel approaches to stopword identification in burmese. In *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Ye Kyaw Thu, Khaing Hsu Wai, Thazin Myint Oo, Myint Myint Htay, and Naing Linn Phyo. 2023b. mypoetry. `https://github.com/ye-kyaw-thu/myPoetry`.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Khaing Hsu Wai, Ye Kyaw Thu, Swe Zin Moe, Hnin Aye Thant, and Thepchai Supnithi. 2020. Myanmar (burmese) string similarity measures based on phoneme similarity. *Journal of Intelligent Informatics and Smart Technology*, 1(1):27–34.

Wint Theingi Zaw, Shwe Sin Moe, Ye Kyaw Thu, and Nyein Nyein Oo. 2020. Applying weighted finite state transducers and ripple down rules for myanmar name romanization. In *Proceedings of the 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2020)*, pages 143–148, Virtual Conference Hosted by College of Computing, Prince of Songkla University, Thailand.