MYANMAR ENTITY IDENTIFICATION FOR NATURAL LANGUAGE UNDERSTANDING USING BIDIRECTIONAL LONG SHORT TERM MEMORY (BiLSTM)

SAUNG THAZIN PHWAY

M.C.Sc. **JANUARY 2023**

MYANMAR ENTITY IDENTIFICATION FOR NATURAL LANGUAGE UNDERSTANDING USING BIDIRECTIONAL LONG SHORT TERM MEMORY (BiLSTM)

By

SAUNG THAZIN PHWAY

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

January 2023

ACKNOWLEDGMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin,** Rector, the University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win and Dr. Tin Zar Thaw,** Professors, Faculty of Computer Science of the University of Computer Studies, Yangon, for their superior suggestions, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr. WIN Pa Pa,** Professor, Department of Computer Science , the University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I also wish to express my deepest gratitude to **Daw Aye Aye Khine**, Associate Professor, Department of English, the University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

Last but not least, I especially thank my parents, all of my colleagues, and friends for their encouragement and help during the period of my thesis.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…….. ……

Date Saung Thazin Phway

ABSTRACT

Entity identification is an exacting function which has commonly appropriate broad chunk of awareness in the course of feature engineering and word list to attain great achievement.. Entity Identification (EI) is indispensable of perceptive article character from basic input and resolve the division the morphemes characterizes. This paper presents every Entities Recognition (ER) for Myanmar language using Bidirectional Long Short Term Memory (BiLSTM), eliminating the need for most feature construction. Entity contains people, location, grouping, date_time_month, numerical values, etc. Myanmar expression is still ambitious to analyze Name Entity (NE) as well as familiar conversation so it bags of geographical instruction towards noticeable items, never barrier explanation among words and none capitalization comparable other languages. Myanmar Natural Language Processing (NLP) is told to be closed growing along with has directly been excruciating to be matured. Considering that logic, Entity Identification (EI) entitled collection for Burma ER analysis is annually explained and built as composing that monograph. The elucidate EI bulk is crucial for Myanmar ER research's improvement . For planned entity classification research, those entity titled compilation is tested all the while entire the aimed evidence for Burma ER and it will also be determined. By using BiLSTM based network architecture, the best accuracy is achieved with 83.62%. Accordingly, here task dispose of the aspect engineering development and does not demand to acquire not only expression but also territory ability.

Keywords: bilstm, deep learning, myanmar entity identification

CONTENTS

		Page
ACKNOWLEDGMENTS		**i**
ABSTRACT		**iii**
CONTENTS		**iv**
LIST OF FIGURES		**vi**
LIST OF TABLES		**vii**
LIST OF EQUATIONS		**viii**
CHAPTER 1 INTRODUCTION		**1**
1.1	Objectives	1
1.2	Organization of the Thesis	2
CHAPTER 2 BACKGROUND THEORY		**3**
2.1	Related Work	3
2.2	Related Components	4
2.2.1	Entity Identification	4
2.2.2	Language Aspect	4
2.2.3	Domain Element	4
2.2.4	Entity Part	5
2.3	Approaches to Names and Entity Recognition	6
2.3.1	Dictionary Lookup Based NER	6
2.3.2	Rule-based NER	6
2.3.3	Statistical-based NER	7
2.3.4	Deep Learning Approach to NER	7
2.4	Some Earlier Exploration on Myanmar NER	8
CHAPTER 3 MYANMAR ENTITY IDENTIFICATION USING BiLSTM		**9**
3.1	Summary of Myanmar Language Style	9
3.2	Myanmar Words	10
3.2.1	Types of Myanmar Entity	11
3.3	Neural Networks For Deep Learning	11
3.3.1	Neural Networks Settlement	11
3.4	Dispersed Representation	11

3.4.1 Word Embedding 12
3.5 Bidirectional LSTM 13
3.6 Data Preparation 14
3.6.1 Defined Entity Types 15
3.7 System Flow of Myanmar Entity Identification Model 17
CHAPTER 4 SYSTEM DESIGN AND IMPLEMENTATION 19
4.1 Neural Modeling for Myanmar Entity Recognition 19
4.1.1 Experimental Setup 19
4.1.2 Pretrained Embeddings 20
4.1.3 Hyperparameters Tuning 20
4.2 Experimental Data Setting 21
4.3 Evaluation Matrix 22
4.4 Confusion Matrix for Myanmar Entity Identification 23
CHAPTER 5 CONCLUSION, LIMITATIONS AND FURTHER EXTENSIONS 27
5.1 Advantages 27
5.2 Restrictions and Farther Extensions 29
AUTHOR'S PUBLICATIONS 30
REFERENCES 31

LIST OF FIGURES

Figure		Page
Figure 3.1	Difference between CBOW and Skip-gram	19
Figure 3.2	Bidirectional LSTM Model	20
Figure 3.3	System Flow of Myanmar Entity Identification	23
Figure 3.3	The System Flow of Myanmar Entity Identification	21
Figure 4.1	Training For Myanmar Entity Identification	26
Figure 4.2	Testing For Myanmar Entity Identification	27
Figure 4.3	Formulas for Evaluation Matrices	28
Figure 4.4	Evaluation for Myanmar Entity Identification	29
Figure 4.5	Confusion Matrix of Myanmar Entity Identification	30
Figure 4.6	User Interface Food Tag for Myanmar Entity Identification	31
Figure 4.7	User Interface QTY Tag for Myanmar Entity Identification	32
Figure 4.6	User Interface ORG for Myanmar Entity Identification	31

LIST OF TABLES

Table		Page
Table 3.1	Example Sentences from Myanmar Entity Tagged Corpus	20
Table 3.2	Lists of Entity Type	20
Table 3.3	Corpus Data Statistics	21
Table 4.1	Hyperparameter Values	25
Table 4.2	Data Statistic Table	27
Table 4.3	Measures of F1 score for Myanmar Entity Identification	30

LIST OF EQUATIONS

Equation 4.1……... 13
Equation 4.2……. 13
Equation 4.3……. 13

CHAPTER 1

INTRODUCTION

The project of illuminating and dividing the items from the transparent content in to pretend article groups is Entity Identification (EI). Especially, it desires to observe contention that are being applied like categories in a accustomed text and determine these identified items into singular brands of entities. A fundamental part is Natural Language Processing (NLP) systems for automated questionings and solutions systems, knowledge improvement, association derivation, characterization, analogue, record management system, automated indicator,search engines, etc. Alterations and uncertainly of items forms including frequent arguments are formal topics. When the EI problem responses, proper arguments are usually experienced. Those issues and the expression behavior, characteristics, designs, realm, and brands, etc., entire have touched on category recognition's achievement. NLP can means documents from syntax, arrangement, error, attitude. It will be Natural Language Understanding (NLU) which will benefit the machine infer the intent behind the language text.

In knowledge expression, EI is an important factor and fundamental effort. EI has collected the subtend consideration in late lifetime. The EI for Burma language style is actually needed to text-file style clarification for Myanmar-language style. Entity-annotated corpus for Myanmar-language are bounded and distributed. The major sense is Myanmar NLP decreased after correlated to another. Annually arranged and entity-annotated corpus for Myanmar-language is suggested to indicate reserve constraint issue. Presently, there are exactly up 3375 sentences and up 18740 entities in that annually entity-annotated corpus. The advantages of deep learning on Burma Language deplores that statistics are focused employing BiLSTM network. BiLSTM is applied to neural networks to Myanmar EI in this thesis

1.1 Objectives

Myanmar NLP is placed at growing case by the time related in that of another regions and each one state has been accomplishing to grow their language-style methods. It is assumed which that task will be convenient in advancement of Myanmar-NLP exploration task. The main objective is to support a high kind EI model for Myanmar-language. Furthermore, that is contracted to indicate source constraint complication into Myanmar-

language counting so one of the major limit to grow NLP analysis is source failure.

The another open-minded of this thesis are as pursues:

- Figure accessible entity-tagged corpus for subsequent
- Accommodated the demands to source glitch entity-recognition in Myanmar
- Provide a high condition Myanmar Language for EI model
- Pare the use of overpriced of increased features-engineering
- Interject a method to undoubtedly indue entities
- Learn the advantages of Myanmar-entity on deep learning
- Develop EI tools for Myanmar language

1.2 The Institution of Thesis

The thesis is coordinated in to five chapters. In Chapter (1), the system's introduction, objectives of the thesis are described. In Chapter (2), Background theory, relevant task and their similar theory are explained. The Chapter (3)Myanmar Entity Identification Model using BiLSTM . In Chapter (4), the design and implementation of the Myanmar-Entity Identification system is expressed. In the final Chapter (5), the conclusion, advantages and restrictions, and farther extensions of the system are conferred.

CHAPTER 2
BACKGROUND THEORY

There is no another applicable corpus which has as more text like that annually assemble entity-tagged corpus for Myanmar-language. Growing entity-tagged corpus is imperative and it is so indispensable as ER modeling for Myanmar. The dataset form ALT corpus was used.

2.1 Related Work

Many researches created for entity-identified in English and another language. For the thesis described in, it will be an attention on the classification of Myanmar-name as it is just now closed in Burma NLP. They aimed to support encoding partial lexicon tests a novel method in neural network. That model develops the performance of CoNLL2003 and Onto Notes datasets. Two lexicons use constructing from publicly available on resources, 91.62 on CoNLL-2003 of F-measure and outcomes 86.28 on Onto Notes of F-measure.

Huanzhong Duan and Yan Zheng, [3] presented that feature templates varies in window size and the fitting of sequence level are so essential for Chinese names entities. The contribution of F score measure predicted for CRF by adding Chinese characters, part of speech, prefix and suffix. The results show that choosing proper factor templates and succession the sets of label may develop the accuracy of CNER, abbreviate the process of model training and pare the consumption of system source.

Hesheng Xu and Bin Hu [4] proposed LSTM-CRF deep learning Model. The experiments issues showing the model's F-measure tracked over the sequence-labeling word-corpus is 88.13% . For the two forms of entities,names of location and institution, the collecting F-measures with the Bi-LSTM-CRF model applying word-segmentation are percent of 67.60 and 89.45, properly better than the measure of F-score attained by applying character-segmentation model. Accordingly, that model implying word-segmentation is more adequate for identifying continued categories.

In Z. Huang, W. Xu, and K. Yu. [5], researched that the based models for sequence tagging using all kinds of LSTM. By comprising the networks of LSTM and BiLSTM, LSTM - CRF and BiLSTM-CRF, sequence tagging is constructed. At first,NLP criterion sequence tagging data set is used by BiLSTM-CRF model that can construct closing to performance on

POS, extraction and NER datasets. In sum, that is vigorous and has less vulnerability in embedding of word as matched to before observations CRF on NLP criterion the tagging of sequence on data sets. The model BiLSTM-CRF could construct closing on POS by chunking and datasets of NER. In sum, this is vigorous and low vulnerability on word-embedding to matched on earliest experiments.

2.2 Related Components

ER was inclined consideration with the association research along with a work, any linked parts arose up which have to be studied although employing in that range. In the midst of those functions, the factors are so essential language, territory or document file and types of entity which are being discovered.

2.2.1 Entity Identification

Entity Identification indicates to the extraction of data work which is acceptable of discovering entities across the text to classify the items of name exists. The Sixth Message Understanding Conference (MUC-6) evolved the term “Named Entity”. MUC was exciting over Information Extraction (IE) works and people get informed the need to divide the information units alike including people, institution and place names, and number expressions that means date-time-month and percent explanations. Those items classifies getting a main duty of IE and ER .

It was created want to form more EI research with a large focus on information extraction. Entity identification trains on user search quires for utilizing entities which spams of texts are entities. Since entity identification and classification analysis had been out stood applying information engineering among machine translation methods and it had been increased with constant experiments and large scientific practice.

2.2.2 Language Aspect

One of the most powerful factors in making thesis is EI. For definite language built previous attacks over ER so that was hopeless changing those systems to numerous languages comfortably. In the other hands, it was achievable to choose characters autonomous can change from language to language with machine learning. This language concerned the work of ER system with substantially.

Concerning the English language, it has led the majority of EI research. Self-determination and multilayered language come the core objection in this area. Since MUC-6

conference ,attention had given Japanese and German were well designed in prior studies and CONLL-2003. Moreover, Dutch and Spanish concerned language conference on CONLL-2002 imposed. French, Italian and Greek have been rangely considered and Chinese has been examined in a generous article. Many another languages such as Korea, Hinduism, Bulgaria, Poland, Roman, Russia, Sweden and Turki became consideration too. Thailand, Indonesian, Malaysian, Vietnam, India and Laos, have been getting in decades and analysis on those literature are in process. Arabian and Mongolia have made many analysis and Myanmar ER came to have debate and still above the before phase.

2.2.3 Domain Element

Another dominant case in EI is domain accommodation. In sum, the arts and type of the context corpus that are competent on can well influence the state of EI. Only elemental matter of arts humiliation ,the state and the text structure are lightly distinct. Little thesis is chiefly lighted to vary realm and text structures. Those analysis says that inspite of acceptable provide to any language, sorting the only domain corpus with other just exists like a large chance.

The developing flow of generating user text on social has treated a demand for EI to justify to this riotous text structures. The state-of-the-art EI on social platform among waits after well adjusted the structure of text.

2.2.4 Entity Part

The types of entity which are being considered is other main view. Formally, there are three types of named-entities were considered which were generally called the annotated scheme. Those three types of named entities are names of “people”, “institution”, and “place”. Therefore, there were four types of entity type “MISC” was describe for those names in CONLL-2 and CONLL-3. Person, Geographical and Political Entities (GPE), Organization and Facility were defined as tags-set for data at Automatic Content Extraction ACE-2003, ACE-2004 and ACE-2005 that are four distinctive entity types. Other two more types Vehicles and Weapons were added in this corpus. The studied types in language are “People”, “Institution”, “Place”, “Number”, “Date_time”, and “Miscellaneous”. A number of 150 entity types are broadly classified and the more items are affirmed to identify to be harder.

2.3 Approaches to Names and Entity Recognition

Either linguistic based or statistical (machine learning) techniques are methods of EI other approaches, the hybrid method is the adding two approaches. The calculational analysis at naturally classifying entities in documents had been began after on last decades. Researchers tried to improve state of EI with large technologies and so it comes a numerous and undefined categories pooling of approaches, methods and demonstrations. A detailed analysis of algorithm which calculates on interested and national disciplines describes to extract and identified entities. ER had been the main of consideration and different learning approaches had accomplished more computational thesis. Suggestion with linguistic approach, dictionary look-up and rule-based approaches are the starting tasks. But lately statistical machine translation sequence-modeling approach was used and combined two approaches was also tested. Different deep neural architecture eliminates feature engineering in recent studies.

2.3.1 Dictionary lookup based NER

List of names (gazetteers) to classify the occupancy of names in text, always by various sub-string matching techniques use utilizing Dictionary lookup based approaches. Only a complete list of names is built that is the belief and names in a described text are found up in the resounding the lists of name. Essentially, name dictionary or thesaurus and Word-Net are desired.

Dictionary based approach is smooth, quick and more acute, and it can donate very well precision comparing with other approaches. Moreover, selection and arrangement of name dictionary is so overpriced and annoying. Fathermore, it is difficult to cap all variants in name dictionary so more and more entities are becoming continually and it could not solve immobility. NER is not still comparing strings with attentively built lists of entities but one identifies entities that are being applied as entities in a disposed text. After, dictionary based technique do not use individually, but that is often applied as function in another methods. A disadvantage of this method is the want of building and persevering the sources.

2.3.2 Rule-based NER

A set of rules is annually organized by linguists. It expresses common on syntactic, linguistic and domain knowledge to identify a singular entity type. This rules are training and the output is given by comparing the rules. While lexical and syntactic suggestions are often applied to make rules. Those rules are not acceptable to classify all experiments

of entities; there are build upon many certain problems and most rules are expulsion. It is fine that rule-based approach build upon on a rule set and it acquires large experience and grammatical information to determine rules.

The absence of entity-annotated corpus sources for Malaysia language that can be applied to a training data, a rule-based method that makes through three steps was suggested. At the first step, tokenization splits the given sentences into tokens. The second step is part of speech tagging process. The final step, considering a tokenism word is whether or not one of the three types of NE (person, location, and organization) was generally established on the POS rule-based tagging process and textual rules.

2.3.3 Statistical-based NER

The statistical approach is very diverse from rule-based approach where discovering and identifying the entity completely commit on the rules described with the linguists; it applies arithmetical formulas and logic to seek and identify entity. Statistical models are naturally manufactured from linguistically annotated-source. In statistical approach, training module is run to recognize and the occurrence of named entities in the corpus, a feasibility is computed. When a context is paid at any time, it fixed on the expectation value.

During training process, distinctive features and some mathematical implementations through the system learn to classify entities creates using handcrafted rules so statistical approach alters from early approaches. Those are three objective learning methodologies to study the system: supervised, semi-supervised learning and unsupervised learning. Feature selection must account critical process and development of annotated corpus when running with statistical approach. The performance of NER and errors in annotated training corpus affects selection of features badly to machine learning based models. The opportunity of huge and tolerable datasets annotations gets the main cons when employing through with a statistical method.

There are distinctive models feeding from Decision Tree to Conditional Random Fields (CRF) and all those models have their individual arithmetic approaches for training and deciding the probabilistic values in growing statistical approach. In addition, each model has inherent methodologies of running to attain employing with statistical approach,the fitting result. In the midst of more flavor models, Hidden Markov Model (HMM), Maximum Entropy, support vector machine (SVM), and Conditional Random Fields (CRFs). By combination with different machine learning approaches is too applied.

2.3.4 Deep Learning Approach to NER

The performance absolutely build upon the analysis of the thesaurus. As like, it is acquired to have linguistic information to fit disciplines for rule-based approach. Moreover, statistical methods to NER are more powerful than rule based approaches. And, these methods desires indigenous features and a huge set of linguistic knowledge to classify Entity adequately. In summation, these approaches are well build upon the selection of appearance. Particularly, deep learning has been broadly used to sequence-tagging in many languages, and a shift of focus from feature-engineering to design and implement the active deep neural network architectures.

Neural network models are accomplish of moderating the concern of statistical model that the need works with arranged appearances, like deep layers on neural networks could study relevant task features naturally. EI system exacting on multitask and multilingual link studying was suggested by unit upon characters adds word embeddings. Those embeddings were passed through another Recurrent Neural Network (RNN) layer and the generated output was given to CRF models employed for disparate works like POS, chunking and ER. ER with neural networks was studied to indicate entity shortage dataset with limited human annotation. Deep Learning improves the state-of- the-art impacts and might be chiefly helpful for a object dataset with deal number of labeling of sequence.

2.4 Some Earlier Exploration on Myanmar ER

There are only two earlier experiments on Myanmar ER. Earlier pursuits upon the work of NER for Myanmar language had been made by focusing classic approaches. The work for Myanmar Named Entity Identification had presented hybrid method. That approach is a consolidation of ruled-based and statistical N-grams based approaches that use name database. They identified Myanmar NEs into three classes, namely person_name (PER), organization_name (ORG) and location_name (LOC). They had inspected a example of 43 Myanmar-text documents. Their analysis paid percent of 82.75 on precision and percent 83.47 on recall above the given data. In the algorithm, the system determines the names by applying the POS knowledge,entity identification disciplines and solution words in the left or the right texts of NEs perform knowledge for NE classification. So, input layer of sentence must be segmented with fashioned POS tags and achievement exactly build upon the linguistic rules. Furthermore, there is a deficiency that is the uncertainly of sematic assumption on suitable names. Linguistic knowledge and extraneous features are elemental

for Myanmar NER.

CHAPTER 3
MYANMAR ENTITY IDENTIFICATION USING BiLSTM

This chapter has expressed the nature of Burma language and the launch of Myanmar Entity Identification. Moreover, chances experienced in clarification entity identification for Burma language have also been explained. This chapter also represents the approaches and nature of deep learning methodologies and generally used distinctive neural networks in NLP, and recent tendency of deep learning in NLP are shown in detail.

3.1 Summary of Myanmar Language Style

The official language of the Republic Union of Myanmar is historically also known as Burmese also called Myanmar language. A cordial of tonal language is used by more than 50 million people. In Burma language, the least linguistic unit is syllable and single word exists singular or other syllables.

Myanmar language is well rational with none inflation of characters which aids that morphemes can be mixed freedom with never changes. The functional dependent morphemes succeeding contented in dependence characters is commonly head-final and the verb capacity by running the root of a sentence is usually at the last of a sentence. Before their flexible parts and the main clause of a sentence places accessory clauses. Generally, there is no definite rule or assembly on the apply of spaces to break words in Myanmar, and spaces are actually used in Myanmar texts illogically to annotated substantial essence.

3.2 Myanmar Words

Generally, Myanmar writing script includes absolutely 75 characters. These characters can be more classified into 12 groups such as 34 consonants, 4 medial letters, 8 dependent vowels, one Sign Virama and one Sign Asat [1]. One group determines three independent vowels, three independent various signs and the characters in these group can perform by standing alone syllables and 10 Myanmar digits and 2 punctuation marks. Other group is hammered with four independent vowels and one Myanmar Symbol preceding. By adding to those characters, white space is applied between phrases and also no obvious rule to apply it. Myanmar-Numerals are decimal-based, and displays zero to nine in series. Thousand of separators are not used; in spite of, spaces are frequently applied among digits for clear reading documents. The punctuation marks part in a related style to the coma (,)

and the duration in other languages like English, respectively[1].

3.2.1 Type of Myanmar Entity

A person name generally exists a family name and a correlated name in a naming system of another countries. Moreover, a type of naming system is not used in most Myanmar names. Basically, general titles could be applied with a person name and determined function of name affixes. For example, it is very familiar for person names to be predated by some various of title in language. Like, in Myanmar language, generational titles predated the person name, e.g., မောင်မောင်မမ etc., although some writings can ignore those titles. Many name appends can be applied with people names, that support knowledge about the people, indicate that the particular controls environment, education, diploma, station, or dignity. These are always not determined as functions of entity,although they support samples of extraneous documentation in acceptance names or entities in text.

These features plays the start of the name such as General, Professor, and Doctor, etc in Myanmar names. So it is possible to be a person name if an expression is predated with alike personal affixes. Although, names infrequently apply or not a predating expressions word to show a person's position, old_age and gender-specific in Myanmar calligraphy. In text, names of persons may be in a subject, object, possessive and comparison place and equal in unity. Words works from left to right and spaces are added to spread clauses kinds of words. Entity Recognition for Myanmar language is a starting state. Other, ER for Myanmar language is told to be very hard correlated to another languages so it is complicated futures.

3.3 Neural Networks For Deep Learning

A neural network on a convinced matched of complication is a seep neural network. The first approach were trivial that consists of one input, output and at one or more hidden layer in between. At least three layers is including input and output that authorized for deep learning. So Deep neural networks imply worldly arithmetical sequence modeling to perform input data in convoluted methods.

3.3.1 Neural Network Settlement

Deep neural networks could be hard to settle. These are more attitude which could comfort to develop the system. The architecture might study gently and the score might not be delighted, if the architecture hyperparameters are badly selected or as it may be not at all.

The major iterance among value of parameters are studied by the model during the training time and hyperparameters while hyperparameters could be altered before running the model. A deep neural network' s parameters are the Weight and Bias, that the architecture changes all along the back breeding stride. In the other, these are many values of hyperparameters for a neural network, includes

- Input normalizing
- Weight initiation
- Training estimate – *a*
- No of repetition (Epoch)
- No of concealed layers
- Entity in each concealed layer
- Renewing function
- Defeat function
- Formulation (e.g. Dropout, Early stopping and weight decay)
- Mini batch size
- Selection of developing algorithm, and so on.

3.4 Dispersed Representation

For modeling complex NLP tasks has been elementary preferred. Moreover, that analytical NLP gets from the notable the expletive of dimensions case although training link feasibility parts on network. This heads to the aim of training assigned presentations of texts in little dimension space that main idea is to present contents like selection-features vectors. The important advantages of the flat presentation is memorization strength that can support a presentation which is capable to collect relationships. A dispersed presentation is possible only the main developments for the important accomplishment in achievement of learning approaches on energetic NLP issues.

3.4.1 Word Embedding

A kind of word representation where separated words are presented as real value vectors in a pretend direction area. Word embeddings are much practiced like the first input data preprocessing layer in to deep learning architecture. Word embeddings are generally run by developing extra aims into a huge non labeled input; primarily studied pass text which the

studied word directions can collect chiefly grammatical and lingual knowledge. So, those word-embeddings have voiced to be active in automated collection of selected-features from context. The main support for the deep learning models results in state-of-the-art achievement.

In entity recognition, studied dispersed presentations for texts have grown. The task which describes the pretrained service of word embeddings has still been accomplished. The provisional feasibility of a tagged text shown the content texts neighboring it with a aperture size. The neighboring text shown the middle tagged text are generated (see Figure 3.1). The content texts are affected to be positioned consistently in to the tagged texts with a length balances to the aperture size into two directions.

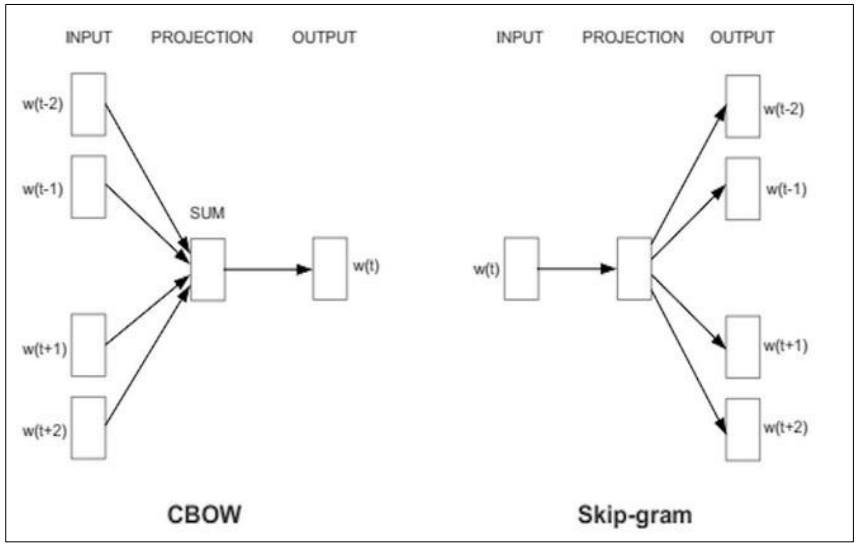

Figure 3.1: Different between Skip-gram and CBOW (Figure Source: [3])

3.5 Bidirectional Long-Short Term Memory

The basic training unit for label tagging treats as annotation by giving a Myanmar sentence. For sequence leaning, the input representation of each input sequence with word break and learned representation was fed into BiLSTM network. In implementation of model training flexible choices of feature inputs and output structure, we employed the production of PyTorch Framework. Model training will work on GoogleColab on Jupyter Note Book. As to optimization, the BiLSTM (Adam) was execute 0.0001 for initial learning rate and set as 64 for batch sizes. The performance on validation sets uses early stopping . In the whole experiment set as 128 for hidden dimension. For BiLSTM,on character information of sequences from both directions is memorized. By concatenating the two hidden states, the final

output is established. This advantage of maintaining information for long times representation experiments happens 30 epochs the accuracy.

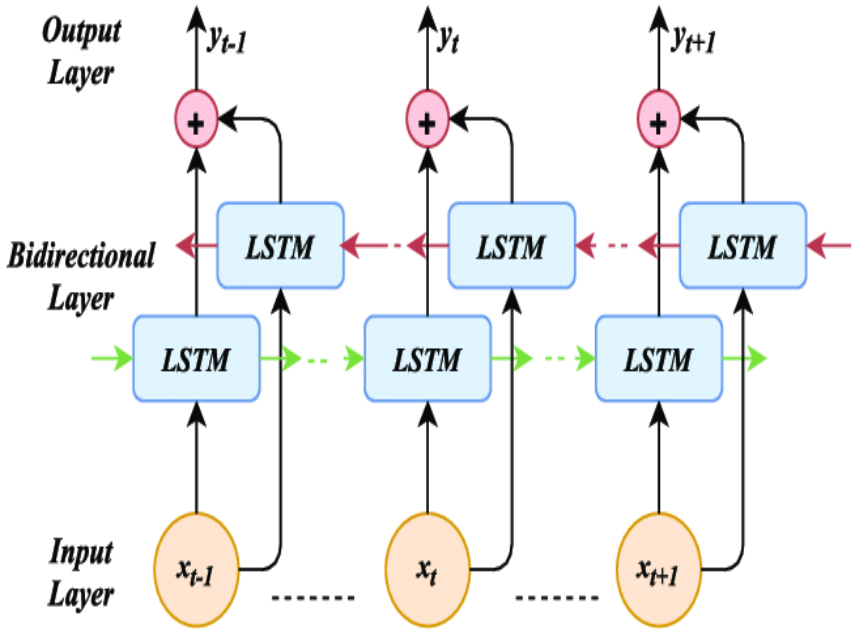

Figure 3.2: Bidirectional LSTM Model

3.6 Data Preparation

In preparation of data, data cleaning is the first start to performed. Various kinds of missing typing errors are precise annually. Addition, contrasting cipher want to be in a reliable concealing. So, entire the captured data are changed into standard Unicode encoding for encoding consistency. Thus it is essential to accurate such various kinds of unfairly typing errors so the condition of raw data greatly influences the achievement.

3.6.1 Defined Entity Types

To express entities in sentences, each entity has to be segmented with Entity tag. In this work, totally sixteen types of entity tags are decided for manual annotation: 'PER', 'LOC', 'ORG', 'DTM', 'TIME','' FOOD', 'PRON', 'N', 'V', 'O', 'CUR', 'TITLE', 'FW', 'PUNC' ,'NUM'. 'PER' tag is used to announce person names while 'LOC' tag is defined for location entities. In this case, politically or geographically defined places are considered as location entities. In addition, location entities include man-made structures like airports, highways, streets, factories and monuments, etc., ' ORG' tag is defined to annotate names of organizations (government and non-government organizations, institutions, corporations,

companies and other groups of people defined by an established organizational structure). ‘TITLE’ identifies the name of the start (e.g Mrs., Ms., Mr.). FOOD classifies Cuisines, various type of food such as Chinese Food, India food, Myanmar food etc. Number identifies numeric number. QTY recognizes quantity, distance count. DTM identifies date, time, month, and year. N represents a thing. V represents the action or state in a sentence. PRON refers to identify someone. CUR identifies the system of money general use in a particular country. TIME identifies indefinite continued progress of existence or events. FW means foreign word of a country or language. Some example segmented sentences from the Entity-tagged corpus are described in Table 3.1

Table 3.1: Example Sentences from Myanmar Entity Tagged Corpus

၁။ ခင်ဗျား **@PRON|** ရဲ့ **@O|** အာလူးမီးဖုတ် **@FOOD|** မကြာခင် **@TIME|** ရတော့မှာ **@V|** ပါ **@O|** ။ **@PUNC|**
၂။ ကိုကာကိုလာ **@FOOD|** တစ် **@NUM|** ဗူး **@QTY|** နဲ့ **@O|** တခြား **@N|** က **@O|** တော့ **@O|** လိမ္မော်ရည် **@FOOD|** တစ် **@NUM|** ခွက် **@QTY|** လောက် **@O|** ။ **@PUNC|**
၃။ ကျွန်တော့် **@PRON|** ကြက်သားကြော် **@FOOD|** က **@O|** စား **@V|** ဖို့ **@O|** မာကြော **@V|** လွန်း**@O|** တယ် **@O|** ။ **@PUNC|**
၄။ အာလူးမီးဖုတ် **@FOOD** | နဲ့ **@O** | အချဉ်နှစ် **@FOOD** | နဲ့ **@O** | ပေး **@V** | ပါ **@O** | ။ **@PUNC** |
၅။စားပွဲထိုး **@N** | တစ်**@NUM** |ယောက် **@QTY** | က **@O** | နေမကောင်း **@V** | လို့ **@O** | လွန်ခဲ့တဲ့ **@TIME** | တစ်နာရီ **@DTM** | လောက်က **@N** | ပြန်သွားခဲ့လို့ပါ **@V** | ။ **@PUNC** |

Further, the description of defined entity categories and some sample usage of each entity category are shown in Table 5.1.

Table 3.2: Lists of Entity Type

Define Entity types	Description
PER	Person name or family(ဒေးဗစ်စမစ်၊ဝမ်၊လီ)
ORG	governmental or cooperate name(အကအဖွဲ့)
LOC	Location name of publicly or graphically decided location(တရုတ်)
TITLE	မစ္စတာ ၊မစ် ၊ မစ္စ

FOOD	Various kinds of food such as Chinese , India, Myanmar food etc.(ပေါင်မုန့်မီးကင်)
NUM	Numeric numbers(တစ်၊ နှစ်)
QTY	Distance, Money, Quantity, Count(ခု၊မိုင်)
DTM	Time, year, Month, days, and periods(တစ်နှစ်၊ နှစ်နှစ်)
N	Represent a thing(အိတ်)
V	Action or state in a sentence(သွား၊ စား)
O	Undefined categories (ဒါ့အပြင်)
PRON	A way to identify or refer a someone(ကျွန်တော်)
CUR	money in typically apply in a country(ဘတ်၊ ဒေါ်လာ)
TIME	Indefinite continued progress of existence and events.(လွန်ခဲ့တဲ့၊အခု)
FW	Not Myanmar words(natural)
PUNC	"။"

In Table 3.2 the entities distributes in manually segmented entity tagged corpus. The occurrence of each entity-tagged corpus is also shown in percentage. It is observed that pronoun entity is the most appear type in the entity annotated corpus.

Table 3.3: Data Statistics Corpus

Data	Total No of Entities	Occurrence of one entity (%) in Entity Tagged Corpus
Sentence line number of entity	31630	26187
PER	210	26
ORG	75	9
LOC	2091	89
Title	60	8
FOOD	736	94
NUM	635	40
QTY	383	45
DTM	1069	39
N	3510	219
V	4459	268
O	8581	620
PRON	1135	112

CUR	207	30
TIME	174	15
FW	505	6
PUNC	3357	255

The findings of one decided entity type in the segmented-tagged corpus is also described in Table 3.3. Entire defined Entity types, the TITLE type is the most occurred entity in the annotated-corpus that is up to 13% out of full Entities. The entity of ORG tag is the least appeared entity tag that is less than 2%.

3.7 System flow of Myanmar Entity Identification Model

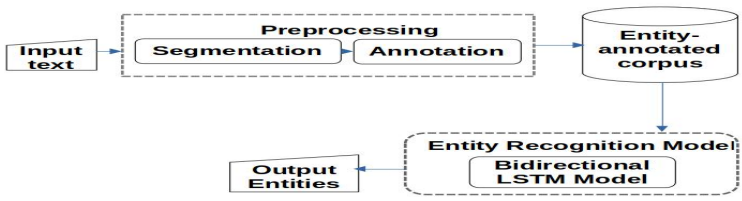

Figure 3.3: System Architecture of Myanmar Entity Identification

In training Phase, input text is segmented or annotated in preprocessing step. And then Entity-annotated corpus from annotated entity is trained Entity Identification Model using BiLSTM. In testing phase, Entity segmented corpus from segmented data is developed using BiLSTM is generated output entities. Complicated sentences and a vary of dissection rich language has Burma language. In the other, well known preparation language sources acquired for Myanmar NLP analysis have not been acceptable still today. Because of agglutinative languages, Myanmar has convoluted linguistic structures;so the architecture can go through from input data fault. Another for architecture in that texts are determined as elemental entity

to built appropriated presentation, these might apparently be complications of these affluent linguistics texts. In the other, drained vocabulary words and word-segmentation issues are crucial statements wanted to carry out all along legitimate statement solution for Burma language. Word-segmentation is essential to preprocess of the largest text-file clarification for the sense which ordinary spacing of white are not added between texts in reported Myanmar sentence contents. It defines which annotation impact will alter the matched of items identification achievements when texts are created like essential entity for appropriated presentation.

Requiring no task for specific resources models, feature-engineering, or data preprocessing above pretend affricate amending on unlivable corpus. Also, the model can be clearly used to a wide broad area of arrangement works on contrasting realm.

CHAPTER 4
SYSTEM DESIGN AND IMPLEMENTATION

In Chapter 4, the improvement of the entity-tagged corpus for Burma language was represented too. That entity-annotated corpus is constructed to resolve source failure. Experiments will be discussed and evaluation of entity identification Model for Myanmar sentence would be represented by implementation.

4.1 Myanmar Entity Recognition Model

Definite commentary of the architecture for Myanmar entity model heaving and developing fit up stairwell are defined in coming areas. So that the entity identification model is available to be employed in the implementation of entity recognition system.Myanmar entity identification Model returns contents from Asian Language Treebank (ALT) CORPUS from NLP Lab).The input contents are divided into word break to train to bidirectional LSTM model to predict output as a result.

4.1.1 Experimental Build up

To appliance the neural network architecture in developing, the neurological libraries supported with the PyTorch framework are practiced so it supports malleable selections of selected-feature inputs and output sentences. The developments are trained on Google Colab on JupyterNotebook. The sentence of input layer into the model divides words level presentations of word embedding. Adam algorithm is tried. The initial training rate was seeeds 0.0001, Batch sizes was seeds 24 for the Adam algorithm, All along developing, established on the achievement on recognition seeds, previously blocking was applied in order that it could delay as well the finest dropout was also used all along the developing. It is seen that dropout developing is important for well abstraction achievement and encouraging upon setting up the developing action. The invisible length was seeded into 128 in the entire development.

In addition, that developing model for Myanmar entity identification is inclined a Myanmar sentences. Word break is checked as the basic developing entity for sequence label tagging. The input presentation of character arrangement is essentially studied by using BiLSTM model within each segmented tagged input.

4.1.2 Pretend Embedding

The association which might independently be so varied, arrive into ordinary contents in huge corpus. Although, embeddings studied from a huge corporations which are conscious to word-order are applied. The pretend word-embeddings are tested to start lookup table. Words embeddings are pretend applying a variation of word2vec and skip-n-grams which counts for word-order. Word-embeddings are developed applying the 3375 sentences from ALT-parallel-corpus data fine-tuned embeddings. 128 of installing proportion, 4 for a minimum word frequency and 8 of window size are applied. Although, pretend installing could not allow finest solutions for various noise data in training during the analysis.

4.1.3 Hyperparameters Tuning

The choice of hyperparameters shows a crucial part in neural network developing. The essential of distant network design selections hyperparameters calculated for morphologic arrangement annotated-tagging tasks. Including training, dropout , no of layers, hidden size, and could powerfully infects the network architecture achievement. The analysis are focused by controlling hyperparameters setting. In that section, hyperparameters used training for Myanmar entity identification will be shown in neural networks. Table 4.1 outlines the hyperparameters allow the finest achievement all along the training.

Table 4.1: Hyper parameter values

Hyper parameters	Value
Learning rate	0.0001
Learning decay	0.05
Hidden dimension	128
Character hidden dimension	128
Average batch size	24
LSTM Layer	2

Epochs	30

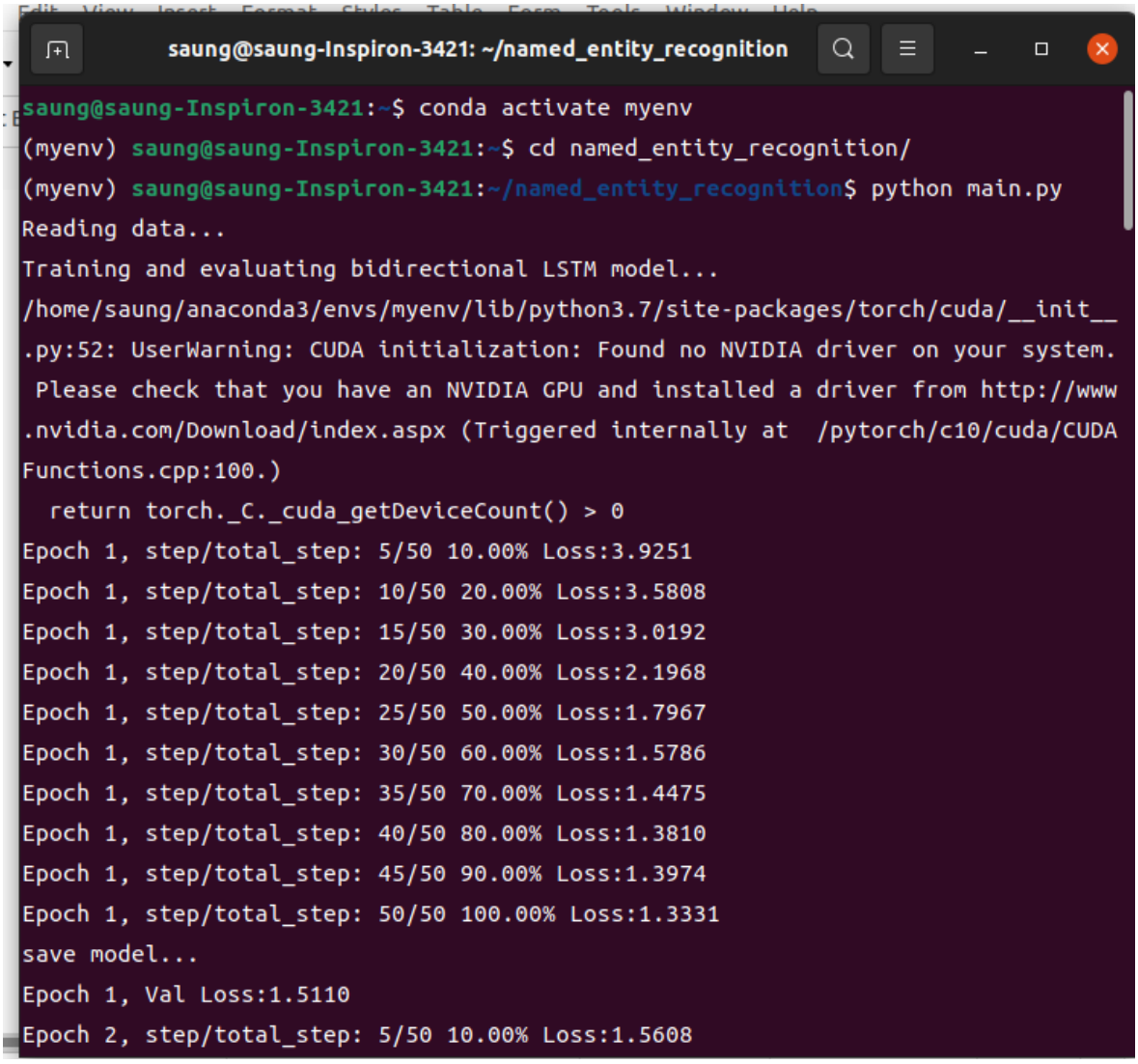

Figure 4.1: Training For Myanmar Entity Identification

4.2 Experimental Data Setting

The system will accept a sentence entity annotated-tagged sets. The entity annotated-tagged values the present and environments tokens as selected-features will be applied. That model will decide pronoun, verb and other categories in a document to classify items with their applicable categories as People, Organization, Title, Food, Number, Quantity, Currency, Date_time_month, Punctuation, etc. If the system found Noun, Verb, Other tag with their concerned classes by applying the arrange rules.

In entity-tagged corpora for Myanmar, these are absolutely sentences 3375 and overall number of entities are 17413. These input data are spread into three tagged sets, for training

(Train), development (Dev) and testing (Test). Overall entity number in individual set is described as data statistics in Table 4.2. In development set, it has only 2% of total entities. In testing phase shows in Figure 4.2.

Table 4.2: Data Statistic Table

	Number of sentences
Training	31630
Validation	3604
Testing	2128

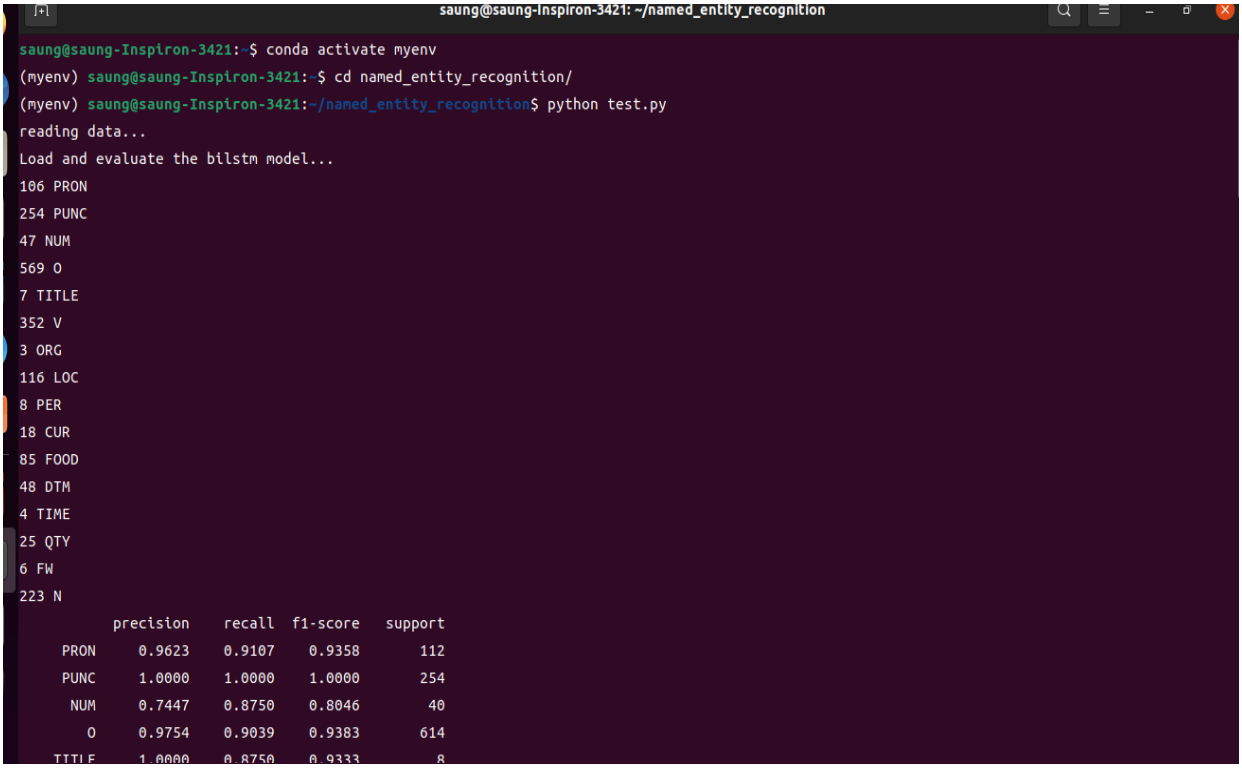

Figure 4.2: Testing For Myanmar Entity Identification

4.3 Evaluation Matric

Evaluation matric is important to classify and analyze outcomes of approaches. ER is usually calculated with general calculation assessment, that applied three metrics known as precision, recall, and F-score to determine the achievement of ER architecture. Precision is a calculation which suggests the portion of the copied entities which are accurate, where Recall is the portion of the perfect entities which are cited. In other hands, Precision can be evaluated

by classifying the no of item which a model generated precisely by the number of items that the model generated. Recall can be evaluated as the number of items that a model generated rightly split by the number of categories which are classified by the human annotations. The F-measure is a symphonic among Precision and Recall and can be told that its attitude is to accurate an total accuracy. Evaluation matric of Myanmar Entity Identification Using BiLSTM shows in Figure 4.3.

$$\text{Precision} = \frac{\text{Number of accurately extracted entities}}{\text{Number of extracted entities}} \qquad \text{equation(4.1)}$$

$$\text{Recall} = \frac{\text{Number of accurately extracted entities}}{\text{Number of all entities}} \qquad \text{equation(4.2)}$$

$$\text{F-measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{precision}+\text{Recall}} \qquad \text{equation(4.3)}$$

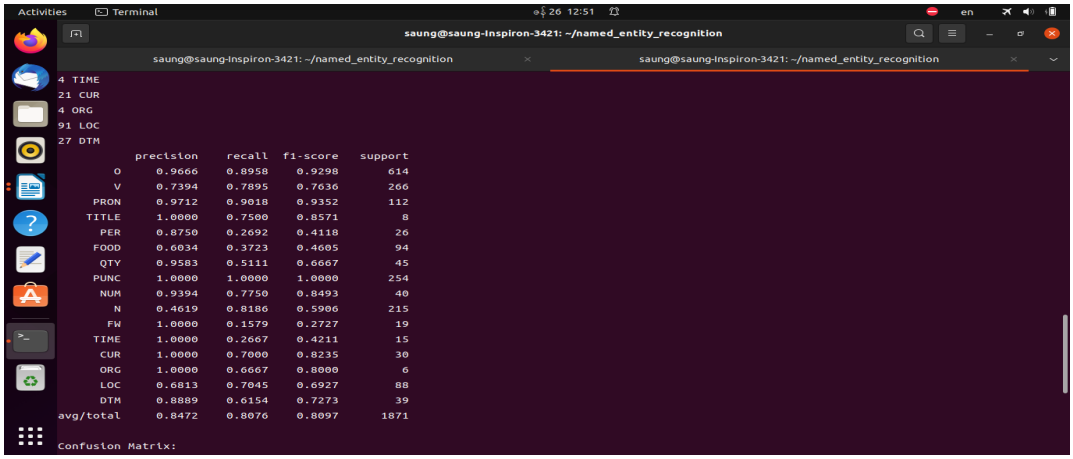

Figure 4.3: Precision, Recall and F1-score of each entity for Myanmar Entity Identification

4.4 Confusion Matrix for Myanmar Entity Identification

To define the achievement of a allocation algorithm and entire the oblique elements stand for corruptly confidential issues. The misconceived issues are presented on the off oblique of the confusion matrix. Therefore, the finest identifier will include a confusion matrix generated actual ideals and generated issues after the clarification action. Figure 4.5 shows the

confusion matrix Myanmar Entity Recognition using BiLSTM .

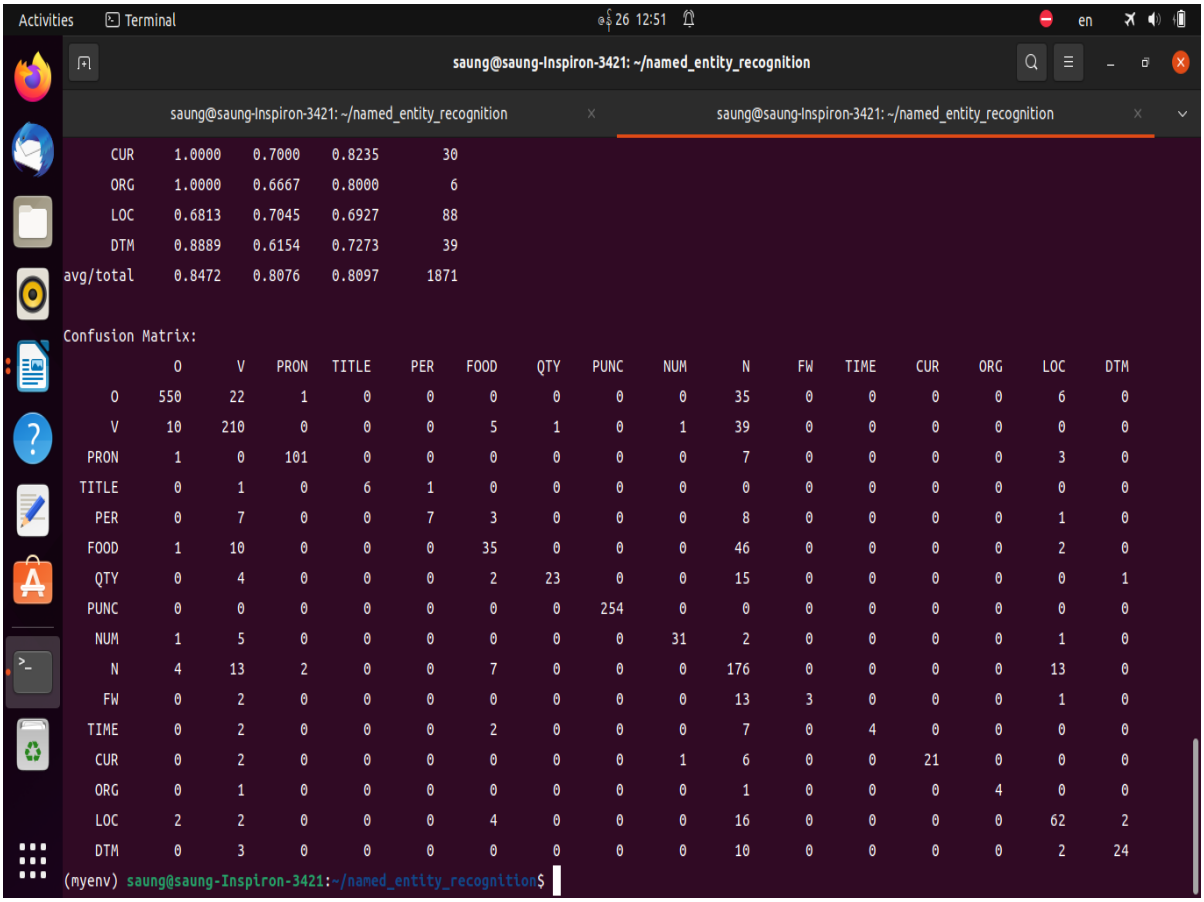

Figure 4.5: Confusion Matrix of Myanmar Entity Identification

Table 4.3: Total Measure of Precision ,Recall , F1 score for Myanmar Entity Identification

Model	Precision	Recall	F-score
BiLSTM	83.64	82.52	82.35

User Interface requires Myanmar sentence in “ပေါင်မုန့်မီးကင်” annotates for “FOOD” tag, “နှစ်” annotates for “NUM” tag, “ချပ်” annotates for “QTY”, "ပေး” annotates “V” , "ပါ” annotates “O” , “။” identifies “PUNC” for Entity Identification Model using BiLSTM that it is generated output entities as shown in Figure 4.6.

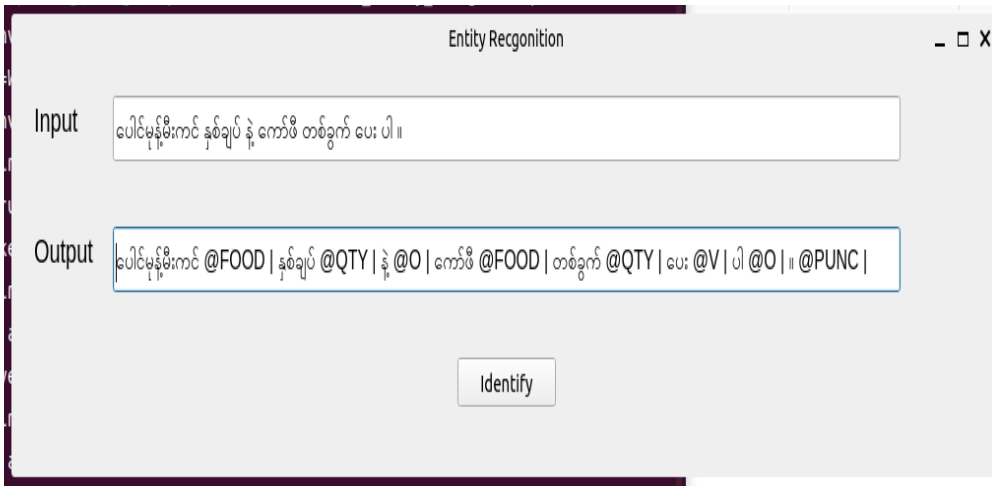

Figure 4.6: User Interface for FOOD Tag for Myanmar Entity Identification

User interface for input sentence is “ကျွန်တော့် ကြက်သားကြော် က စား ဖို့ မာကြောလွန်း တယ် ။"The annotated output sentence shows “ကျွန်တော်" identifies “PRON”, "ကြက်သား ကြော်" identifies “FOOD” , “က" recognizes “O”, "စား" identifies “V”, "ဖို့" identifies “O”, “မာကြောလွန်း" identifies “V” , "တယ်" recognizes “O” , “။" recognizes “PUNC” for identification model using BiLSTM that is shown in Figure 4.7.

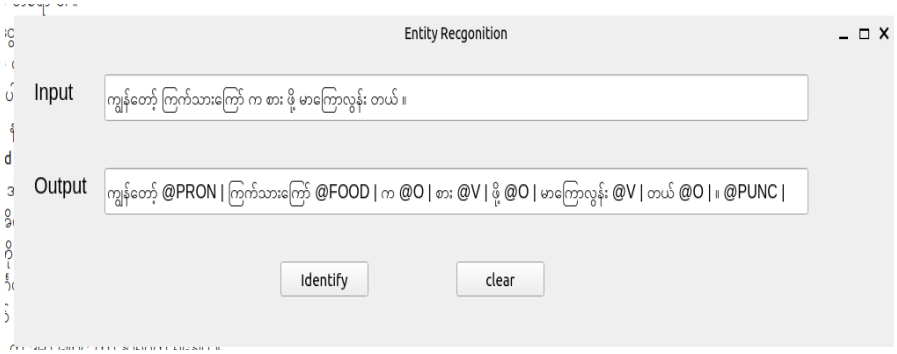

Figure 4.7: User Interface in Food Entity tag for Myanmar Entity Identification

User interface to identify LOC tag for input sentence is “တရုတ် ဆပ်ကပ်အဖွဲ့ က ဖျော်ဖြေ မှာ ပါ ။" The output sentence with “တရုတ်" classifies “LOC”, "ဆပ်ကပ်အဖွဲ့" classifies “ORG” ,”က" classifies “O”, “ဖျော်ဖြေ" classifies “V” , "မှာ" identifies “O”, "ပါ" identifies “O” , “။" identifies “PUNC” using Myanmar Entity Identification using BiLSTM as shown in Figure 4.8.

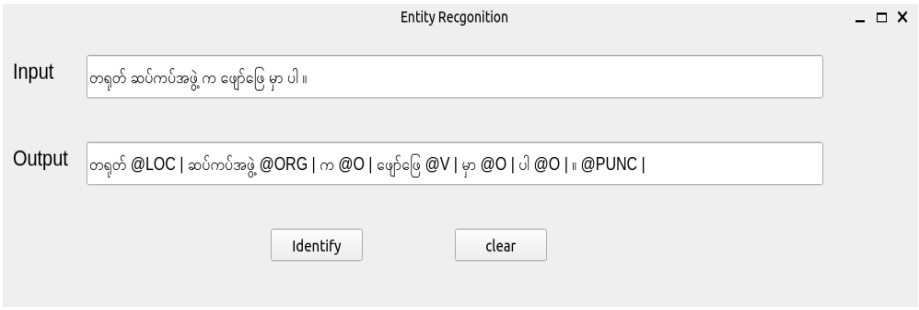

Figure 4.8: User Interface in ORG Tag for Myanmar Entity Identification

CHAPTER 5

CONCLUSIONS, LIMITATIONS AND FURTHER EXTENSIONS

This chapter is the detail of explanation of the thesis work, as well as the improvement and condition of the thesis. The central addition of the thesis task is the so early assessment of neural network architecture on Myanmar Entity Recognition. As part of this thesis, Entity-tagged corpus for Myanmar language is manually advanced. Words are appropriated as input tokens in neural ER modeling.

5.1 Advantages

Entity Identification benefits to explain the architecture of content and to asset communications among entities. In sum, EI can be get to develop the certainly of another NLP research, like as part-of-speech tagging and dependency parsing. Natural Language Understanding (NLU) is a subliming of unreal agility that facilitate the analysis of data by reasoning it, applies it into computer language and generating an output in an unsustainable design for humans. Entity identification for NLU grants user to collaborate with the computer employing common contents.

5.2 Limitations and Further Extensions

Myanmar ER brings improved achievement after any further selected-features. It completes improved than conventional statistical models as it has been disclosed in earlier chapter.

It is able to classify entities in daily conversation style sentences. It could also identified person who are not anticipate by title words and location externally any indication texts neighboring around them. It can be told that this neural model has the ability to identified and recognized into predesign entity correctly excepting some basic errors. Many entities could also be identified by that model. For TIME type, there are two forms:

numerical style or test style. Different TIME tags are accurately accepted. This neural model for Myanmar ER can be combined into the improvement of Myanmar ER tool, Information Retrieval system, entity linking, etc.

Furthermore, improvement of annually entity-annotated tagged corpora can commit in eventual thesis on Myanmar ER. It can assert in improvement of Myanmar NLP research work. Correctly, there are sixteen defined Entity tags in this Entity corpora. From this annotated-entity tagged corpora, entities can be chunked and entity lists can also be built if acquired

In the time, there are just disadvantages and limitations in this neural entity classification for Myanmar language. As limitation, it can only means on contents that are recorded in Unicode encoding. Entities written in addition of English and Myanmar Characters cannot be identified so developing data are not carried out for such problems. Entity in sentences which are written in message form are not adequately identified as good. There are many ambitious fault in that model.

A lately direction in Deep Learning is accelerating on consideration machine sets. Consideration mechanism has got acclaim freshly in image, speech and NLP fields for entity identification appreciate their model performings. In eventual, based attention neural analysis will be controlled with the motive of advance achievement of Myanmar ER.

Data in the annually annotated entity tagged corpus is not as much as another languages so that more and more data wants to be added as more as possible. In eventual, Entity corpus will be built with many more defined Entity tags and even in stratified contents.

Although the Entity tagged corpus is not too huge, neural network models project finer achievements models for Myanmar Entity Identification, we still accept with more input data and more analysis, forward looking neural networks could study higher because of production of better results. Besides, there is an intention to do it realm autonomous.

With more test data and more analysis, higher results will be recorded in the eventually and deep neural networks will be stored exploiting on Myanmar Entity identification and also on other Myanmar NLP research, e.g., POS tagging and word segmentation, too. Moreover, in the eventually, Myanmar ER system is pretend to construct by end-to-end learning approach.

AUTHOR'S PUBLICATIONS

[1] Saung Thazin Phway, Win Pa Pa, "*Myanmar Entity Identification for Natural Language Understanding Using BiLSTM*", The Proceedings of the Conference on Parallel & Soft Computing (PSC 2022), University of Computer Studies, Yangon, Myanmar, 2022.

REFERENCES

[1] H.M.MO, and K.M.Soe. “Syllable -Based Neural Named Entity Recognition For Myanmar Language “, International Journal on Natural Language Computing (IJNLC) vol.8, No.1,February 2019

[2] P.C.J.Chiu ,and E. Nichols, “Named Entity Recognition with Bidirectional LSTM-CNNs”, Transactions of the Association for Computational Linguistics , vol. 4 (2016):pp.357-370,2016

[3] Huanzhong Dun and Yan Zheng, “A Study on Features of the CRFs-based Chinese Named Entity Recognition”, International Journal of Advanced Intelligence. Volume 3, Number 2, July, 2011, pp.287- 294.

[4] Hesheng Xu and Bin Hu“Legal text recognition with LSTM-CRF deep learning Model”, vol. 2022, March 17, 2022.

[5] Z. Huang, W. Xu, and K. Yu. “Bidirectional LSTM-CRF Models for Sequence Tagging,” vol.1508.01991. 2015 Aug 9.

[6] R. Alfred, L. C. Leong, C. K. On, P. Anthony, T. S. Fun, M. N. B. Razali and M. H. A. Hijazi, “A Rule-Based Named-Entity Recognition for Malay Articles,” AD MA 2013, Part 1, LNAI 8346, Springer-Verlag Berlin Heidelberg 2013, pp.288-299.

[7] M. Ali, G. Tan, and A. Hussain. "Bidirectional Recurrent Neural Network Approach for Arabic Named Entity Recognition," Future Internet 10, no. 12 (2018): 123.

[8] K.S. Bajwa and A. Kaur, “Hybrid Approach for Named Entity Recognition,” International Journal of Computer Applications, vol.118, no.1, pp-0975- 8887, May. 2015.

[9] P. Basile, G. Semeraro, and P. Cassotti, "Bi-directional LSTM-CNNs-CRF for Italian Sequence Labeling," CLiC-it 2017 11-12 December 2017, Rome (2017).

[10] I. E. Bazi, and N. Laachfoubi, "A Comparative Study of Named Entity Recognition for Arabic using Ensemble Learning Approaches," In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), pp. 1-6. IEEE, 2015.

[11] Y. Benajiba, M. Diab and P. Rosso, “Arabic Named Entity Recognition: An SVM-based Approach,” In Proceedings of 2008 Arab International Conference on Information Technology (ACIT) 2008 (pp. 16-18). Amman, Jordan: Association of Arab Universities.

[12] Y. Bengio, R. Ducharme R, P. Vincent, and C. Jauvin, “A Neural Probabilistic Language Model,” Journal of Machine Learning Research, 2003;3(Feb):1137-55.

[13] D. Bonadiman, A. Severyn, and A. Moschitti, "Deep Neural Networks for Named Entity Recognition in Italian," CLiC it 51 (2015).

[14] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition," In Sixth Workshop on Very Large Corpora. 1998.

[15] R. Caruana, S. Lawrence, CL. Giles, “Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping,” In Advances in Neural Information Processing ySstems 2001 (pp. 402-408).

[16] H.L. Chieu, and H.T. Ng, “Named Entity Recognition with a Maximum Entropy Approach,” In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 160-163). Association for Computational Linguistics, 2003, May.

[17] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, “On theProperties of Neural Machine Translation: Encoder-decoder Approaches,” arXiv preprint arXiv:1409.1259. 2014 Sep 3.

[18] J.K. Chorowski, D. Bahdanau, D. Serdyuk D, K. Cho, Y. Bengio Y, “Attention-

based Models for Speech Recognition," In Advances in Neural Information Processing Systems 2015. (pp. 577-585).

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv preprint arXiv:1412.3555, 2014.

[19] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and Pavel Kuksa, "Natural Language Processing (almost) from Scratch," Journal of machine learning research 12, no. Aug (2011): 2493-2537.

[20] A. Das and U. Garain. "CRF-based Named Entity Recognition@ ICON 2013," [Online] Available: https://arxiv.org/abs/1409.8008.

[21] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks," [Online] Available: https://arxiv.org/abs/1705.05487, May, 2017.

[22] C. Ding, Y. K. Thu, M. Utiyama, and E. Sumita, "Word Segmentation for Burmese (Myanmar)," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol.15, no. 4 (2016): 22, May 2016.

[23] A. Ekbal, and S. Bandyopadhyay, "A Hidden Markov Model based Named Entity Recognition System: Bengali and Hindi as Case Studies," In International Conference on Pattern Recognition and Machine Intelligence, pp. 545-552. Springer, Berlin, Heidelberg, 2007.