

# Name Entity Recognition for Myanmar Language

Thida Myint

University of Computer Studies, Mandalay

thidamyintucsm@gmail.com

## Abstract

*Name Entity Recognition (NER) is one of the fundamental problems in many Natural Language Processing applications. NE consists of names, person, location, organization, time, numerical expressions and etc. Not much works have been done on NER for Myanmar language. Due to the characteristics of Myanmar language, NE extraction is more difficult than English Language. Myanmar Language is also difficult to distinguish NE and common word because it lacks of orthographical information to signal NEs, no boundary indication between words and no capitalization like English. In this paper, we propose NER algorithm which can recognize three classes of NEs: Person, Organization and Location. In identification phase, we employed pre-defined rules generated from Language knowledge, and Myanmar-English bi-lingual Lexicon.*

## 1. Introduction

Named Entity Recognition (NER) is an important tool in almost all Natural Language Processing (NLP) application areas. Proper identification and classification of named entities are very crucial and pose a very big challenge to the NLP researchers. The level of ambiguity in named entity recognition (NER) makes it difficult to attain human performance.

NER is a task that seeks to locate and classify entities in a text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, etc. It can be viewed as a two stage process:

- (i) Identification of entity boundaries
- (ii) Classification into the correct category

For example, if “**ဦးမြင့်ဆွေ**” is a named entity in the Person Name, it is necessary to identify the beginning and the end of this entity in the sentence. Following this step, the entity must be classified into the predefined category, which is Named Entity Person in this case.

NE is one of the fundamental problems in many Natural Language Processing applications, such as information extraction, machine translation, automatic abstract and data mining. In Machine Translation (MT), NEs require different techniques of translation than the rest of words of the text. Also, the post-editing step is more expensive when the errors of a MT system are mainly in NEs translation. For these reasons, [1] have carried out a research study where they tag a text with a NER system as a pre-processing step of MT.

There is no concept of capitalization in Myanmar language (ML) like English and this fact makes the NER task more difficult and challenging in ML. There has been very little work in the area of NER in Myanmar language. In Myanmar language, the work in NER can be found in [8]. This system accepts the Myanmar sentence as input. This system is based a hybrid

method that is a combination of rules based and statistical N-Grams based approach. In this paper, we accept the Myanmar input sentence with the POS tag set. And then, we identify and classify the NEs with their related classes such as person, organization and location by using the proposed algorithm.

Myanmar does not use space or special character as a delimiter between words. A large portion of Myanmar NE doesn't have certain structure, and can construct any word. There is also a problem from Myanmar writing style by using different form(s) and/or without clue word. This could cause the ambiguity between NE and common word. Even when named entities are listed in the dictionaries, it is not always easy to decide their sense. There can be semantic ambiguities.

In this paper, we report a named entity recognition system for ML. Myanmar language being a member of the Tibeto-Burman language, which is a subfamily of the Sino-Tibetan family of language, is the official language of Myanmar. It is spoken by 32 million people as a first language. The system considers linguistic features. Linguistic features of Myanmar include the set of known suffixes that may appear with named entities, clue words that help in predicating the location and organization names, words that help to recognize measurement expressions, designation words that help in identifying person names, the various gazetteer lists like the first names, middle names, last names, location names and organization names. As part of linguistic features for Myanmar, the system uses only the lists of first names, middle names and last names along with the list of words that helps to recognize measurements.

The rest of the paper is organized as follow. We present related work on name entity class recognition in Section 2. Also a brief survey on nested NE recognition system is presented here.

The challenges in NER are described in Section 3. The classes of NER are described in Section 4. We present the proposed Myanmar NER system and algorithm in section 5 and 6. Finally Section 7 concludes the paper.

## 2. Related Work

Many researchers have been work for name entity recognition in English Language. For the research reported in this paper, we will emphasis on the identification of Myanmar names because it is still now open in Myanmar NLP. In the following paragraphs, we discuss briefly some of the related work and history in the area of Name Entity Recognition.

Huanzhong Duan et.al [3] studied on features of Chinese Name Entity Recognition (CNER) based on Conditional Random Fields (CRFs). They aimed to provid a reference for selecting features of CNER through a series of experiments. The experiment results show that appropriate features or their combinations, such as single Chinese character, part-of-speech (POS), prefix and suffix can contribute to the score of F-measure based on CRFs.

Vijay Sundar Ram et.al [9] presented a study of the impact of different types of language modeling by selecting different feature matrices in the Conditional Random Fields (CRFs) learning algorithm for Named Entity tagging. They had come up with four different feature matrices and identified features at word, phrase and sentence level. It was identified that the language model which has the structural feature is better than the models with other features [9].

R.C. Balabantaray [6] proposed a method for name entity class recognition based on such rules and some assumptions. They evaluated the performance of their method with various kinds of texts. The work presented in their paper is part of a larger effort to develop Machine Translation

(MT) system which can take care of name entities.

Vijayakrishna R et.al [10] presented a domain focused Tamil Named Entity Recognizer for tourism domain. This method takes care of morphological inflections of named entities (NE). They have experimented building Conditional Random Field (CRF) models by training the noun phrases of the training data and it gives encouraging results.

Kashif Riaz [4] studied the challenges of named entity recognition for resource scarce languages among South Asian languages. Urdu is used as an example language because of its large number of speakers, the only language in the region with Arabic script orthography, and interesting assumptions about its similarity with Hindi.

In this paper, we propose named entity recognition algorithm for ML. We recognize three classes of NEs: Person, Organization and Location by using the proposed algorithm. We employed pre-defined rules generated from Language knowledge, and Myanmar-English bilingual Lexicon for checking noun in the identification phase.

### **3. Challenges in Name Entity Recognition**

Named Entity Recognition was first introduced as part of Message Understanding Conference (MUC-6) in 1995 and a related conference MET-1 in 1996 introduced named entity recognition in non-English text. In spite of the recognized importance of names in applications, most text processing applications such as search systems, spelling checkers, and document management systems, do not treat proper names correctly. This suggests proper names are difficult to identify and interpret in unstructured text. Generally, names can have

innumerable structure in and across languages. Names can overlap with other names and other words. Simple clues like capitalization can be misleading for English and mostly not present in non western languages like Myanmar.

The goal of NER is first to recognize the potential named entities and then resolve the ambiguity in the name. There are two types of ambiguities in names, structural ambiguity and semantic ambiguity. The objective of NER is to identify and classify every word/term in a document into some predefined categories like person name, location name, organization name, miscellaneous name (date, time, percentage and monetary expressions etc.) and “none-of-the-above”. The challenge in detection of NEs is that such expressions are hard to analyze using rule based NLP because they belong to the open class of expressions, i.e., there is an infinite variety and new expressions are constantly being invented.

#### **3.1 Challenges in Myanmar Language**

Named Entity (NE) identification in ML is difficult and challenging as:

- i. Unlike English and most of the European languages, Myanmar lacks capitalization information, which plays a very important role in identifying NEs.
- ii. Myanmar person names are generally found in the dictionary as common nouns with some specific meanings. For example, နုနဲ is a person name and can also be found in the dictionary as a common noun with the meaning ‘jewellery’.
- iii. Myanmar is an inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex wordforms. For example, the person name သိတာ [root] can appear as သိတာမြင့်

[inflection:- မြင့်], သီတာဝင်း [inflection:- ဝင်း], သီတာသိန်း [inflection: -သိန်း], သီတာသိမ့် [inflection:- သိမ့်] etc.

- iv. Myanmar is a relatively free phrase order language. Thus, NEs can appear in any position of the sentence making the NER task more difficult.
- v. Myanmar is a resource- constrained language. The annotated corpus, name dictionaries, good morphological analyzers, POS taggers etc. are not yet available in the required measure.
- vi. Although ML have a very old and rich literary history, technological developments are of recent origin.
- vii. Web sources for name lists are available in English, but such lists are not available in Myanmar. This necessitates the use of transliteration for creating such lists.

### 3.1.1. No Capitalization

Capitalization is an important feature used by the English NER systems to identify the NE. The absence of the lexical features such as capitalization in ML scripts makes it difficult to identify the NE.

### 3.1.2. Agglutinative nature

Agglutinative property means that some additional features can be added to the word to add more complex meaning. Agglutinative languages form sentences by adding a suffix to the root forms of the word.

### 3.1.3. Ambiguity

Ambiguity in proper names is present in South Asian languages as in English. Common nouns can be used as proper names in South

Asian languages. One of the properties of the named entities in this language is the high overlap between common names and proper names. Among the named entities themselves, there is ambiguity between a location name ခင်ဦး (သူသည် ခင်ဦးမြို့နယ်တွင် နေထိုင်သည်။) or a person's surname 'မခင်ဦး' (မခင်ဦးသည် ဆရာမဖြစ်သည်။). In this system, we will correctly identify these ambiguity names as person or location if their related clue words are included in the input sentence.

## 4. Classes of Name Entity

The Named entity hierarchy is divided into three major classes; Entity Name, Time and Numerical expressions. The Name hierarchy has eleven attributes. Numeral Expression and time have four and three attributes respectively [7]. There are eleven types of entities in Name as given below.

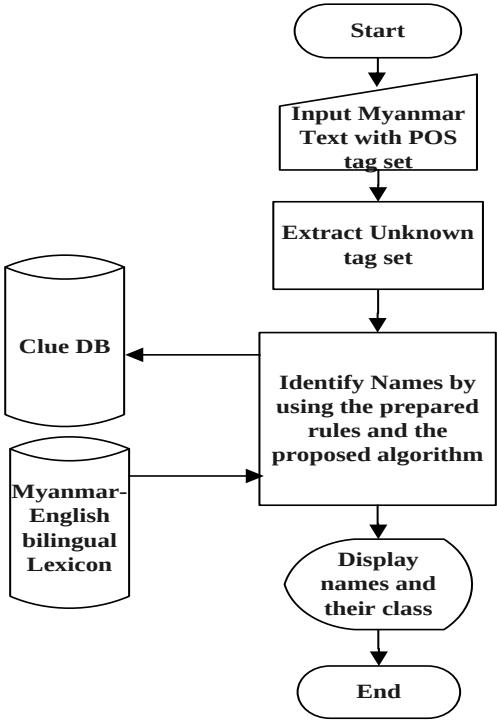
### (a) Name

- (i) Person: named person or family;
- (ii) Organization: named corporate, governmental, or other organizational entity;
- (iii) Location: name of politically or geographically defined location;
- (iv) Facilities: named of buildings and other permanent man-made structures such as Hospitals, Library, Hotel, etc;
- (v) Locomotives: of physical device designed to move an object from one location to another;
- (vi) Artifacts: named of objects or things such as Painting, clothes, Gems & Stone, Medicines, etc;
- (vii) Entertainment: activities such as Dance, Music, Sports, Cinema, etc;
- (viii) Cuisines: various type of food such as Chinese food, India foods

- (ix) Organisms: living things such as viruses, bacteria, etc;
- (x) Plants
- (xi) Disease
- (b) Time: Year, Month, Date, Day, Period;
- (c) NUM: Distance, Money, Quantity, Count;

Although having eleven attributes of name entity, most of researchers are still identified only three attributes (Person, Location, and Origination). In this paper the common three named entities, person name, location name and organizational name are mainly studied.

### 5. Proposed Myanmar Name Entity Recognition System



**Figure1.System architecture for named entity recognition for ML**

The system will accept a sentence that has been tagged with part of speech (POS) tag sets.

We have used the POS values of the current and surrounding tokens as features. This system will search the Unknown Noun (NUNK) tag in a sentence to identify name entities with their related classes such as Person, Location and Origination. If the system found the NUNK tag, we will identify these NUNK tag as names with their related classes by using the prepared rules, the proposed NER algorithm, clue words and bilingual lexicon. This system will continue until not founding the POS tag set in an input sentence.

#### 5.1. Training Features Set

The training feature set of conditional random fields (CRFs) is a very important parameter, which can directly affect the NER results. The main features of CRFs can be divided into two categories: the internal and the external. The former includes character or word, POS, boundary and other context information. Although these basic internal features may have certain effects on the results, the external features are needed to obtain better results. The external features are mainly from the general statistical information, which consists of the prefix of common Myanmar names, the suffix of place names and organizational names and so on. Firstly, the recognition accuracy of person name can be increased by adding the prefix of Myanmar name (mainly common surname or clue words for person in Myanmar, such as ‘ဦး’, ‘ဒေါ်’, ‘ကို’, ‘မ’, etc.). For example, most Myanmar surnames are used as the prefix feature of name, and person name whose length between 2 and 4 is selected as the Myanmar name feature according to the Myanmar named norms. Secondly by adding the suffix of place name or clue words for location (such as ‘တိုင်း’, ‘ပြည်နယ်’, ‘မြို့’, ‘ရွာ’ etc.) to the feature of location recognition, most Myanmar characters are

chosen as the suffix of location by statistics, and the place name whose length not less than two Myanmar characters are also selected as the location formation feature according to the Myanmar place named norms. And then, by adding the suffix of organization name or clue words for organization (such as ‘ကုမ္ပဏီ’, ‘ကုမ္ပဏီလီမိတက်’, ‘ဝန်ကြီးဌာန’, ‘ဌာန’ etc.) to the feature of organization recognition, most Myanmar characters are chosen as the suffix of organization by statistics.

## 6. Myanmar Name Entity Recognition Algorithm

Input: Myanmar sentence with POS tag set  
Output: Names with their classes (Person, Location, Organization)

1. nt = number of tag in an input sentence
2. For (int i=0; i<nt; i++)
3. {
4. If (tag<sub>i</sub> = “NUNK”),
5. {
6. If (tag<sub>i</sub>[0] == clue in Person), tag<sub>i</sub> is Person.
7. Else if (tag<sub>i+1</sub> == human-title), tag<sub>i</sub> is Person.
8. Else if (tag<sub>i+1</sub> == verb OR particle), tag<sub>i</sub> is Person.
9. Else if (tag<sub>i+1</sub> == sentence marker), tag<sub>i</sub> is Person.
10. Else if (tag<sub>i+1</sub> == postposition for Person), tag<sub>i</sub> is Person.
11. Else if (tag<sub>i+1</sub> == postposition for Location), tag<sub>i</sub> is Location.
12. Else if (tag<sub>i+1</sub> == Conjunction for Person), tag<sub>i</sub> is Person.
13. Else if (tag<sub>i+1</sub> == Conjunction for Location), tag<sub>i</sub> is Location.
14. Else if (tag<sub>i+1</sub> == Noun),
15. {
16. If (Noun==features for Organization), tag<sub>i</sub> is Organization.
17. Else if ((Noun==Noun in lexicon) AND (tag<sub>i+2</sub>==features for Organization), tag<sub>i</sub> is Organization.
18. Else if (Noun==features for Location), tag<sub>i</sub> is Location.
19. Else tag<sub>i</sub> is Person.
20. }
21. }
22. }

**Figure2. Myanmar Name Entity Recognition Algorithm**

In this algorithm, the input is a Myanmar sentence with POS tag set. The POS tags set used in this algorithm are described in table 1.

|                           | POS tag                                                      | Myanmar words                                                                                                  |
|---------------------------|--------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| Postposition for Person   | PONOM,POOBJ,POREA,POACC P,POPOS, etc.                        | သည်, ကို, ကြောင့်, အား, ၏                                                                                      |
| Verb                      | VAC, VST, VCP                                                | ပြော, သာယာ, ပြောဆို                                                                                            |
| Noun                      | NUNK,NCCS,NC CP,NCU,NV,NC PS,NCPP,NCPU, NCLS,NCLP,NA BQ,NABS | မောင်မောင်, စာအုပ်, ရွေးများ, ရေ, လူလွယ်အိတ်များ, နားနီ, မိသားစု, မိသားစု, ပရိသေသကများ, တူးချွန်မှု, ချစ်ခြင်း |
| Particle for Person       | PAIDNUM,PAN SA                                               | များ, တို့.                                                                                                    |
| Conjunction for Person    | COW,COCP,CO CB,COCH,COCOR                                    | နှင့်, ထက်, အပြင်--လည်း, သို့မဟုတ်, မှ, သော်လည်းကောင်း--သော်လည်းကောင်း                                         |
| Postposition for Location | PODEP,PODIR,P OARR,POPLA, POT,POPOS,PO COM                   | မှာ, က, သို့, ထိအောင်, ဌာ, ၏, တိုင်အောင်                                                                       |
| Conjunction for Location  | COW,COCP,CO CB,COCH,COCOR                                    | နှင့်, ထက်, လည်း--လည်း, သို့မဟုတ်, မှ, က, ဖြစ်စေ--ဖြစ်စေ                                                       |
| Sentence Marker           | SM                                                           | ,,                                                                                                             |

**Table1. Examples of POS tag set with Myanmar Words**

The algorithm will loop until no other POS tag set in the input sentence. This system will search UNK tag to identify names with their classes such as Person, Location and Organization by using the prepared rules and clue words. If the next tag of UNK tag is Noun, the system will be checked that this is Noun in Myanmar-English bilingual lexicon or features in clue database.

There are a few limitations in this algorithm to identify correctly the ambiguous names for person or location. If the clue words of location are included in an input sentence, the proposed algorithm will be more increase the accuracy of location.

For example, the input sentence is: ရွှေမြန်မာ / UNK / ကုမ္ပဏီ / NCLS / မှ / COCON / ဦးဇော်မောင် /

UNK / သည် / PONOM / ရန်ကုန် / UNK / မြို့ /  
NCCS / တွင် / POPLA / နေထိုင် / VST / သည် /  
POVP / ■ / SM /. Then, the output is: ရွှေမြန်မာ  
[Organization], ဦးဖေမောင် [Person], ရန်ကုန်  
[Location].

The expected outcomes will be texts from Myanmar Grammar book of Myanmar Association, Myanmar newspapers and Myanmar websites by using the Myanmar NER algorithm. These sentences can contain at least one name word and three names words at most.

## 7. Conclusion

NER in Myanmar is a challenging problem for language processing. This system will extract the UNK tag set and identify the boundary of candidate NE that composes of many words by using the pre-defined rules, the proposed NER algorithm and Myanmar-English bi-lingual lexicon. Then, this system will display NEs and correctly classify the class of these NEs. It is important to note that there are so many available features in the Myanmar NER such that the ones concerned in this study are actually not enough for expressing all possible situations of the Myanmar NER. As a preliminary from this study, more investigations are needed in the future work. After that, we want to transliterate these NEs to their relevant Myanmar phonetics. In transliteration, Myanmar syllables are transliterated into English pronunciation.

## References

- [1] Babych B. and Hartley A. 2003. "Improving Machine Translation Quality with Automatic Named Entity Recognition". In Proceeding. of EACL-EAMT. Budapest.
- [2] Berger A and et.al, "A Maximum Entropy Approach to Natural Language Processing", Computational Linguistics, 1996.
- [3] Huanzhong Duan 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] Kashif Riaz, "Rule-based Named Entity Recognition in Urdu", Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 126–135, Uppsala, Sweden, 16 July 2010

[5] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In Proceedings of the IEEE, 77(2):257–286. 1989.

[6] R.C. Balabantaray, "Name Entity Recognition in Machine Translation", International Journal of Computer Science & Emerging Technologies, Volume 1, Issue 3, October 2010

[7] Satpreet Arora, "Named Entity Recognition", Indian Institute of Technology, Bombay

[8] Thi Thi Swe, "A hybrid method for Myanmar named entity identification and transliteration to English", January 2010

[9] Vijay Sundar Ram R and et.al, "Identification of Different Feature Sets for NER tagging using CRFs and its impact", Special Volume: Problems of Parsing in Indian Languages, Language in India www.languageinindia.com, May 2011.

[10] Vijayakrishna R and Sobha L, "Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, Asian Federation of Natural Language Processing, January 2008, pages 59–66.