A Hybrid Method for Myanmar Named Entity Identification and Transliteration to English

Thi Thi Swe, Hla Hla Htay
University of Computer Studies, Yangon
thithiswett@gmail.com, hlahlahtay123@gmail.com

Abstract

Named Entity Identification includes locating named entities and classifying those names in text. NEI is an important task in NLP applications such as Information Extraction, Cross Language information Retrieval, Question Answering, and Machine Translation. In this paper, we presented a method for Myanmar Named Entity Identification using hybrid method. A hybrid method is a combination of ruled based and statistical N-grams based method which use name databases. We have examined a sample of 10 Myanmar text files .We obtained 89% of accuracy in Named Entity Identification. We also classified those named entities into three classes. After that, those names are transliterated to their relevant Myanmar phonetics. We have used transliteration table for mapping with English word. In transliteration table, Myanmar syllables are transliterated into English pronunciation. The system is implemented using Java.

Keywords*:* Named Entity Identification, Ruled Based, Statistical N-Grams Based, Transliteration.

1. Introduction

Named Entity Identification (NEI) also known as Named Entity Recognition (NER) is a key technique in many text-based applications such as question answering, information retrieval, information extraction and machine translation and so on. NEI is relatively simple in English, because mostly English NEs start with an uppercase character. For Myanmar language, no such indicator exists and the problem of identification is more complex, especially for NEs.

There are two main problems in NEI. First, Named Entity (NE) is included in an open word class. Person, location, and organization names can be newly made by the human. Adding such words to dictionary is a very time consuming task and it is impossible to add all NEs to a dictionary. Second, NE can be used as several NE types (ambiguity problem). For example, "Paris" may be used as a person in the sentence "Paris was a prince of Troy." and used as a location in the following sentence "Paris is a capital city of France."

This paper presents a hybrid method that combines two approaches, namely the Ruled Based Approach and Statistical N-Grams Based Approach to identify the NEs. In Ruled Based Approach, NEs are locating by using clue words lists, such left occurrences, right occurrences and left-right co-occurrences. After locating NEs, those are classified according to their clue words. Most of English NER systems used a statistics based approach and in this system used both rules and statistics. Rule based NER system can achieve the proper performance with ease. But, in this system, rules are made by handcrafted. Therefore, it is domain-specific and less portable. This means the rules must be modified according to the domain.

Sometimes, NEs do not exist with clue words. In that situation, we apply NEI by using Statistical N-Grams Based Approach. In second approach, Myanmar texts are firstly segmented into syllables [4]. And then, n-grams syllables of that will also be calculated. After then, the frequency of the unidentified phrases are calculated to identify that phrase is whether NE or not. In this approach, NEs are classified according to their name databases. In this paper, we identify three kinds of Myanmar NEs, namely person name (PER), organization name (ORG) and location name (LOC).

2. Related works

Hau-Ping ZHANG et.al identified Chinese Named Entity using unified statistical model, namely role model. They defined roles as some special token class, including NE components and its neighboring and remote contexts. They also used the Viterbi algorithm to obtain tokens [1].

Hutchatai Chanlekha et.al reported that Thai Named Entity Recognition using statistical and heuristic rule-based model. The idea they used is to make use of small proper name lexicon together with rules, created from internal and external evidence to extract Thai NEs [5].

Alireza Mansouri et.al presented that English NER from text using FSVM for NER to improve the

precision. They had employed Support Vector Machine for classification [2].

Fien De Meulder and Walter Daelemans described a memory-based approach to learning names in English and German newspaper text using memory-based learner Timbl (Daelemans et al., 2002) [3].

Yi-Gyu Hwang et.al presented a named entity recognition model for Korean Language using a HMM based named entity recognition using compound word construction principles [7] .

the right hand side of the clue words and location and organization NEs may find in the left hand side of the clue words. For example, let the phrase ဦးမြတ်မိုးအောင်အား . In this case, if a clue word ဦး find, it will also find the right occurrences like က၊ မှ၊ အား ,etc. After then, the word between such cases may define as a person NEs. Sometimes, if a clue word ဦး , there is no ending clue words. Such case may possible. For example, ဦးမြတ်မိုးအောင် . For LOCATION and ORGANIZAATION NEs, there will not be clue words that occur in the left-hand side. For example,

Table 1. Roles for 3 classes of NEs

Roles	Significance	Examples	Numbers of Clue Words
HP or NI	Head components of Person or Neighboring token in front of NE	ဦး၊ ဒေါ်၊ ကို၊ မောင်.....	8
TP or NF	Tail components of Person or Neighboring token following NE	က၊ မှ၊ အား၊ သည်၊ တို့ကို.....	22
NB	Tokens between two NEs	နှင့်၊ "၊"	2
TO or NF	Tail component of organization or neighboring token following NE	ကုမ္ပဏီ၊ ကုမ္ပဏီမှ၊ ကုမ္ပဏီသို့...	13
NB	Tokens between two NEs	နှင့်၊ "၊"	2
TL or NF	Tail component of location or neighboring token following NE	မြို့နယ်၊ တိုင်း၊ ကျေးရွာ၊ ဒေသ ဒေသ၊ ပြည်နယ်၊ နိုင်ငံ.....	80
NB	Tokens between two NEs	နှင့်၊ "၊"	2

3. Hybrid Method for Myanmar Named Entity Identification and Transliteration to English

Unlike the western language such as English or Spanish, Myanmar NEI is difficult and we cannot straightly borrowed the method the have been used. In English, there are also delimiters and they have the significant definition of each word. In Myanmar, there is no word level segmentation till now.

In this system, we identify Myanmar Named Entity by using hybrid method that combines ruled based approach and statistical n-grams based approach and the resulted Named Entities (NEs) are transliterated according to their relevant phonetics.

Step 1: Ruled Based Approach

In this approach, clue words searching will assign by using the database that are stored separately for the clue word of NEs. That clue words are seen in table1. Named Entity can appear the left hand side or right hand side of the clue words and sometimes they may exist between clue words. Such case will denote the left-right co-occurrence and may mostly occur in PERSON NEs. Normally, person NEs may find in

ရန်ကုန်တိုင်း၊ သုခစုစံကုမ္ပဏီ၊ etc. In ရန်ကုန်တိုင်း၊ သုခစုစံကုမ္ပဏီ , တိုင်း and ကုမ္ပဏီ are clue word of Location and organization.

However, sometimes NEs cannot occur with clue words will be solved with statistical based.

Step 2: Statistical N-Grams Based Approach

In this approach, we use the statistical n-grams based approach. Firstly, we collected *person* names over 10,000 and *location* names over 350. And then, we pre-calculated the frequency of each unigram, bi-gram syllable by position likelihood, such as position1, position2, position1-2, etc.

In this approach, we calculate the frequency of phrases that cannot be unidentified by ruled based approach. Firstly, the phrase is syllabified and each syllable is treated as a gram. We then calculate the unidentified phrase's unigram, bi-gram syllable by position likelihood and their frequencies are calculated by using Name and Location database.

The equation [6] that calculates the frequency is

$$\prod_{i=1}^{i=n-1}\left[\frac{freq\left(pos_i(X_i)\right)}{freq(X_i)} * \frac{freq\left(pos_i(X_i), pos_{i+1}(X_{i+1})\right)}{\left(freq\left(pos_i(X_i)\right) + freq\left(pos_{i+1}(X_{i+1})\right)\right)}\right] \frac{freq\left(pos(X_n)\right)}{freq(X_n)}$$

For example, let the phrase is နွယ်နိုချို့အား . The နွယ်နိုချို့အား equals to 4-syllabled one and its length is 4. In which [နွယ်နိုချို့အား], နွယ်နိုချို့ is NEs and အား is just a particle follow by a NE. In identifying နွယ်နိုချို့ as a NE, two iterations are required.

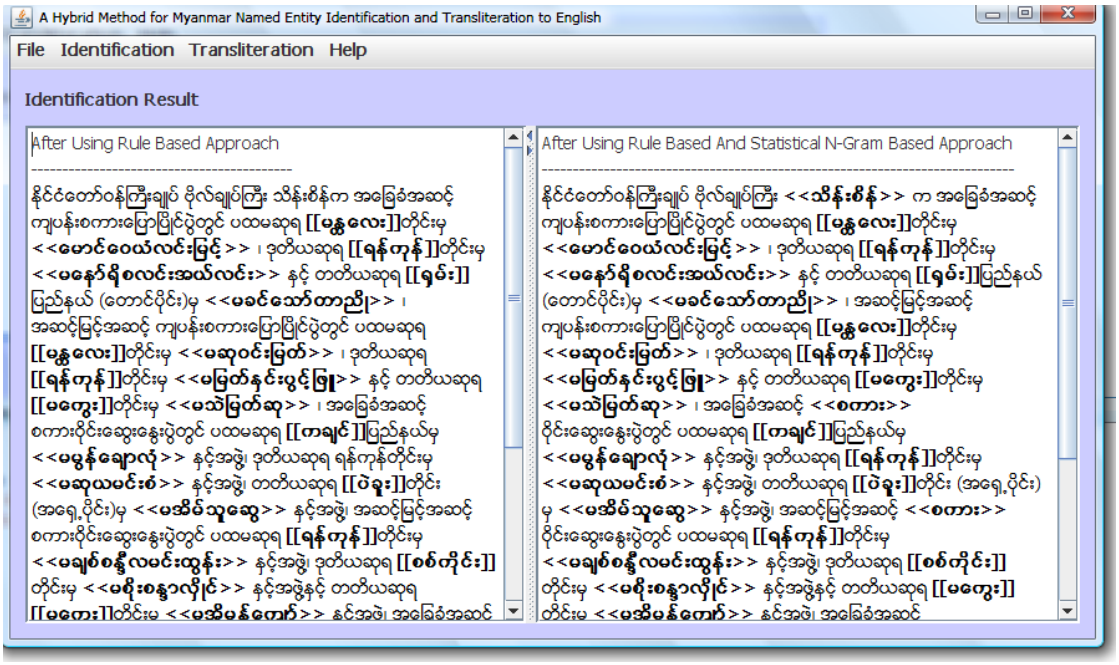

Figure 1. Named Entities identified Myanmar texts

```
for each phrase in the sentence do
  ArraySyllables←Syllabification (Phrase)
  length←ArraySyllables
  while(length>2) do
    for(i=1..length-1)
result=∏_{i=1}^{n-1} [freq(pos1(Xi))/freq(Xi) * freq(pos1(Xi),pos1+1(Xi+1))/(freq(pos1(Xi))+freq(pos1+1(Xi+1)))] freq(pos(Xn))/freq(Xn)
    End for
      if(result==0)
        length←length-1
      else
        Identify that phrase as a NE
        NE found. Match NE.
        Break
      End if
  End while
End for
```

In first iteration, နွယ်နိုချို့အား is calculated as follow.

$$= \left[\left(\frac{freq(pos_1(\text{နွယ်}))}{freq(\text{နွယ်})} * \frac{freq(pos_1(\text{နွယ်}),pos_2(\text{ိ}))}{(freq(pos_1(\text{နွယ်}))+freq(pos_2(\text{ိ}} \right) * \left(\frac{freq(pos_2(\text{ိ}))}{freq(\text{ိ})} * \frac{freq(pos_2(\text{ိ}),pos_3(\text{ချို့}))}{(freq(pos_2(\text{ိ}))+freq(pos_3(\text{ချို့}} \right) * \left(\frac{freq(pos_3(\text{ချို့}))}{freq(\text{ချို့})} * \frac{freq(pos_3(\text{ချို့}),pos_4(\text{အား}))}{(freq(pos_3(\text{ချို့}))+freq(pos_4(\text{အား}} \right) \right] \left(\frac{freq(pos_4(\text{အား}))}{freq(\text{အား})} \right)$$

In above calculation, the frequency result of နွယ်နိုချို့အား is zero because there is no unigram syllable အား in name database and the frequency value of အား is zero. Therefore, နွယ်နိုချို့အား cannot be identified as a NE.

In second iteration, we calculate နွယ်နိုချို့ .

$$= \left[\left(\frac{freq(pos_1(\text{နွယ်}))}{freq(\text{နွယ်})} * \frac{freq(pos_1(\text{နွယ်}),pos_2(\text{ိ}))}{(freq(pos_1(\text{နွယ်}))+freq(pos_2(\text{ိ}} \right) * \left(\frac{freq(pos_2(\text{ိ}))}{freq(\text{ိ})} * \frac{freq(pos_2(\text{ိ}),pos_3(\text{ချို့}))}{(freq(pos_2(\text{ိ}))+freq(pos_3(\text{ချို့}} \right) \right] \left(\frac{freq(pos_3(\text{ချို့}))}{freq(\text{ချို့})} \right)$$

In second iteration, the frequency result is greater than zero. Therefore, that phrase is identified as a NE.

This approach is time consuming if the identified text is so longer. Mostly, Myanmar NEs do not usually start with clue words. Therefore, the second approach is very efficient and useful in identification of NEs that do not start with clue words.

4. Transliteration

Transliteration is the practice of transcribing a word or written in one writing system into another writing system, such that a reader should be able to reconstruct the original spelling of unknown transliterated words. Transliteration is the process of replacing words in source language with their approximate phonetic or spelling equivalents in target language. Commonly, transliteration is used to translate named entity across languages. Although, the bilingual lexicons may be updated from time to time, new named entities out of the lexicons still appear frequently. Automatic Transliteration is helpful for many applications such as Machine

Translation , Cross Language Information Retrieval and Information Extraction , etc.

In this system, we will demonstrate transliteration by using Name Mapping. For NEs နွယ်နီချို , the

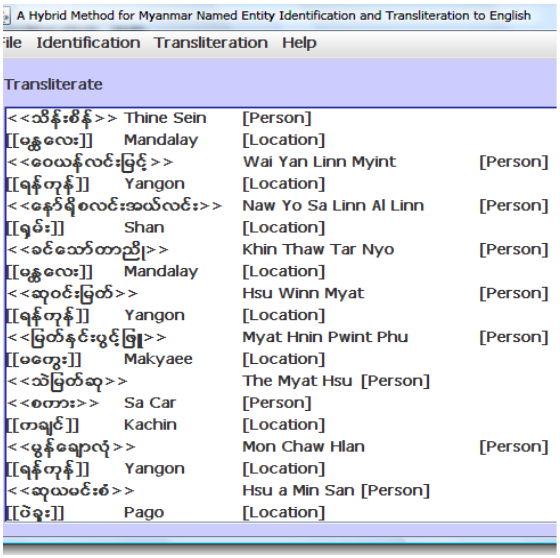

Figure 2. Transliterated Myanmar Named Entities

transliterated NE is Nwe Ne Cho. See details in Figure 2.

5. Error Analysis

We have examined the errors in all three stages of our system.

1. Named Entity Identification: In rule based, the system wrongly identified as named entities if some phrase starts or ends with clue word and in fact, the phrase is open classword. For examples, ဦးဦးဖျားဖျား , မနှစ်မြို့. But this error can be corrected by checking statistical n-grams based approach

2. Classification: A part part of a phrase is misidentified as named entity for example စုစုပေါင်း where the first two syllables are frequently used for ladies' name. This error can be corrected if the dictionary is used.

3. Transliteration: the identified names are needed to improve the phonetic adaptation according to the syllable position although adaption is required 13.5% of the names found.

6. Conclusion

Named Entity Identification includes locating name entities and classifying those names in text. NEI is important task in NLP applications such as Information Extraction, Cross Language information Retrieval, Question Answering, and Machine Translation.

In this paper, we presented a method for Myanmar Named Entity Identification using hybrid method. A hybrid method is a combination of ruled based and statistical N-grams based method. We have examined a sample of 10 Myanmar text files .We obtained 89% of accuracy in Named Entity Identification. We also classified those named entities into three classes. After that, those names are transliterated to their relevant Myanmar phonetics. We have used transliteration table for mapping with English word. In transliteration table, Myanmar syllables are transliterated into English pronunciation. The system is implemented using Java.

We have also discussed the errors and suggested how to tackle those errors. The system can also combine in Myanmar word segmentation process and Myanmar part of speech tagging. All databases can be updated.

References

[1] Hau-Ping ZHANG, Qun LIU, Hong-Kui YU, Xue-Qi CHENG, Shuo BAI, "Chinese Name Entity Recognition Using Role Model " , Computational Linguistics and Chinese Language Processing, Vol. 8,No .2 ,August 2003,pp.

[2] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat, "Named Entity Recognition Using a New Fuzzy Support Vector Machine", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008

[3] Fien De Meulder and Walter Daelemans, "Memory-Based Named Entity Recognition using Unannotated Data", CNTS - Language Technology Group, University of Antwerp

[4] Hla Hla Htay and Kavi Narayana Murthy, Myanmar Word Segmentation using Syllable level Longest Matching, The 6thWorkshop on Asian Language Resources (ALR 6) 11-12 January 2008, Hyderabad, India.

[5] Hutchatai Chanlekha, Asanee Kawtrakul, Patcharee Varasrai and Intiraporn Mulasas, " Statistical and Ruled Based Model for Thai Named Entity Recognition", The Specialty Research Unit of Natural Language Processing and Intelligent Information System Technology, Department of Computer Engineering, Kasetsart.

[6] Paul Wu Horng-Jyr; Na Jin-Cheon; Christopher Khoo Soo-Guan , "A Hybrid approach to fuzzy name search incorporating language-based and text based principles", Journal of Information Science(JIS), http://jis.sagepub.co.uk, Nanyang Technological University, 31 Nanyang Link, Singapore

[7] Yi-Gyu Hwang, Eui-Sok Chung, Soo-jong Lim, "HMM based Korean Named Entity Recognition", Speech/Language Technology Research Center, Electronics and Telecommunications Research Institute, 161, Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea.