International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 4 | May-Jun 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Myanmar Named Entity Recognition with Hidden Markov Model

Khin Khin Lay[1], Aung Cho[2]

[1]Associate Professor, [2]Lecturer
[1]Faculty of Computer Science, [2]Application Department,
[1,2]University of Computer Studies, Maubin, Myanmar

How to cite this paper: Khin Khin Lay | Aung Cho "Myanmar Named Entity Recognition with Hidden Markov Model" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-4, June 2019, pp.1144-1147, URL: https://www.ijtsrd.com/papers/ijtsrd24012.pdf

IJTSRD24012

ABSTRACT

Named Entity Recognition is the process to detect Named Entities (NEs) in a file , document or from a corpus and to categorize them into certain Named entity classes like name of city, State, Country, organization, person, location, sport, river, quantity etc. This paper introduces the Named Entities Recognition (NER) for Myanmar language using Hidden Markov Model (HMM).The main idea behind the use of HMM language independent and we can apply this system for any language domain. The corpus used by our NER system is also not domain specific.

Keywords: *Named Entity Recognition (NER), Natural Language processing (NLP), Hidden Markov Model (HMM)*

I. INTRODUCTION

Named Entity Recognition is a subtask of Information extraction whose aim is to classify text from a document or corpus into some predefined categories like person name(PER), location name(LOC), organization name(ORG), month, date, time etc. And other to the text which is not named entities. NER has many applications in NLP. Some of the applications include machine translation, more accurate internet search engines, automatic indexing of documents, automatic question answering, information retrieval etc. An accurate NER system is needed for these applications. Most NER systems use a rule based approach or statistical machine learning approach or a Combination of these.

A Rule-based NER system uses hand-written rules frame by linguist which are certain language dependent rules that help in the identification of Named Entities in a document. Rule based systems are usually best performing system but suffers some limitation such as language dependent, difficult to adapt changes.

II. RELATED WORK

There are a variety of techniques for NER. NER is classified two approaches:

A. Linguistic approach

The linguistic approach is the classical approach written by linguists to NER. It typically uses rules manually written by linguists. Though it requires a lot of work by domain experts, a NER system based on manual rules may provide very high accuracy. Rule based systems are lexicalized grammar, gazetteer lists, and list of trigger words. The main disadvantages of these rule based techniques are: they require huge experience and grammatical knowledge on the particular language or domain; the development is generally time-consuming and sometimes changes in the system may be hard to accommodate.

B. Machine learning based approach

The recent Machine learning (ML) techniques make use of a large amount of annotated data to acquire high level language knowledge. ML based techniques facilitate the development of recognizers in a short time. Several ML techniques have been successfully used for the NER task. HMM is a ML. Other ML approaches like Support Vector Machine (SVM), Condition Random Field (CRF), Maximum Entropy Markov Model (MEMM) are also used in developing NER systems.

III. OUR PROPOSED METHOD

A. HMM based NER

We are using Hidden Markov Model based machine learning approach. Named Entity Recognition in Myanmar Languages is a current topic of research. The HMM based NER system works in three phases. The first phase is referred to as „Annotation phase" that produces tagged or annotated document from the given raw text, document or corpus. The second phase is referred to as „Training Phase". In this phase, it computes the three parameters of HMM i.e. Start Probability, Emission Probability (B) and the Transition Probability (A). The last phase is the „TESTING Phase". In this phase, user gives certain test sentences to the system, and based on the HMM parameters computed in the previous state, Viterbi algorithm computes the optimal state sequence for the given test sentence.

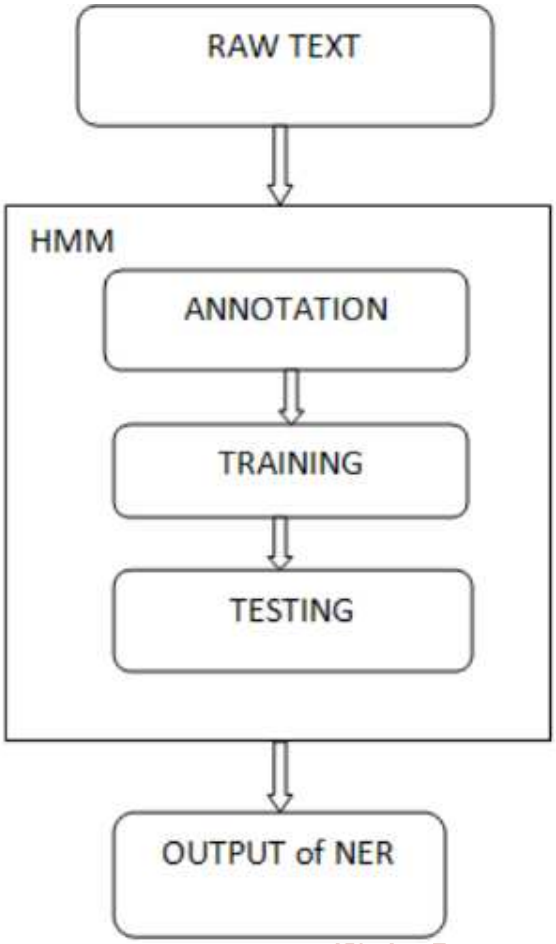

Fig. 3.1: Steps in NER using HMM

B. Step 1: Data Preparation

We need to convert the raw data into trainable form, so as to make it suitable to be used in the Hidden Markov model framework for all the languages. The training data may be collected from any source like from open source, tourism corpus or simply a plaintext file containing some sentences. So in order to make these file in trainable form we have to perform following steps:

Input: Raw text file
Output: Annotated Text (tagged text)

Algorithm
Step1: Separate each word in the sentence.
Step2: Tokenize the words.
Step3: Perform chunking if required.
Step5: Tag (Named Entity tag) the words by using your experience.
Step6: Now the corpus is in trainable form.

C. Step 2: HMM Parameter Estimation

Input: Annotated tagged corpus
Output: HMM parameters

Procedure:
Step1: Find states.
Step2: Calculate Start probability (π).
Step3: Calculate transition probability (A)
Step4: Calculate emission probability (B)

D. Procedure to find states

State is vector contains all the named entity tags candidate interested.

Input: Annotated text file
Output: State Vector

Algorithm:
For each tag in annotated text file
If it is already in state vector
Ignore it
Otherwise
Add to state vector

E. Procedure to find Start probability

Start probability is the probability that the sentence start with particular tag.

So start probabilities (π) =
(Number of sentences start with particular tag) (1.1)
(Total number of sentences in corpus)

Input: Annotated Text file:
Output: Start Probability Vector

Algorithm:
For each starting tag
Find frequency of that tag as starting tag
Calculate π

F. Procedure to find Transition probability

If there is two pair of tags called Ti and Tj then transition probability is the probability of occurring of tag Tj after Ti.

So Transition Probability (A) =
(Total number of sequences from Ti to Tj) (1.2)
(Total number of Ti)

Input: annotated text file
Output: Transition Probability

Algorithm:
For each tag in states (Ti)
For each other tag in states (Tj)
If Ti not equal to Tj
Find frequency of tag sequence Ti Tj i.e. Tj after Ti
Calculate A = frequency (Ti Tj) / frequency (Ti)

G. Procedure to find emission probability

Emissin probability is the probability of assigning particular tag to the word in the corpus or document.

So emission probability (B) =
(Total number of occurrence of word as a tag) (1.3)
(Total occurrence of that tag)

Input: Annotated Text file
Output: Emission Probability matrix

Algorithm:
For each unique word Wi in annotated corpus
Find frequency of word Wi as a particular tag Ti
Divide frequency by frequency of that tag Ti

H. Step 3: Testing

After calculating all these parameters we apply these parameters to Viterbi algorithm and testing sentences as an observation to find named entities. We used the training data 3000 sentences and testing data 150 sentences of Myanmar language.

IV. EXAMPLE

Consider these raw text containing 5 sentences of Myanmar language.

အနောက်ဖက်လွန်မင်းတရားကြီး/PER သည်/OTHER ခဖေဝုဂ်/OTHER ခုနှစ်/OTHER တွင်/OTHER မင်းဆက်/OTHER သစ်/OTHER တစ်/OTHER ခု/OTHER ကို/OTHER အင်းဝ/LOC တွင်/OTHER ထူထောင်/OTHER ခဲ့/OTHER သည်/OTHER ။ /sb

ဘုရားသခင်/OTHER သည်/OTHER မုဟမ္မဒ်/PER အား/OTHER အမိန့်/OTHER ဒေသနာ/OTHER များ/OTHER ကို/OTHER ဟောကြား/OTHER ပို့သ/OTHER ရန်/OTHER စေလွှတ်/OTHER လိုက်/OTHER သည်/OTHER ။/sb

စန္ဒဂုတ္တမောရိယမင်း/PER သည်/OTHER ဂင်္ဂါမြစ်ဝှမ်း/LOC မှ/OTHER တိုင်း/OTHER များ/OTHER ကို/OTHER ပါ/OTHER သိမ်းသွင်း/OTHER ပြီးလျှင်/OTHER ကျယ်ပြန့်/OTHER သော/OTHER မောရိယ နိုင်ငံတော်/LOC ကြီး/OTHER ကို/OTHER ထူထောင်/OTHER နိုင်/OTHER ခဲ့/OTHER သည်/OTHER ။/sb

ကြည်းတပ်/ OTHER ရေတပ်/OTHER နှင့်/OTHER လေတပ်/OTHER တို့/OTHER ကို/OTHER စုပေါင်း/OTHER ၍/OTHER စက်ာပူတပ်မတော်/ORG ဟု/OTHER ခေါ်ဆို/OTHER ကြ/OTHER သည်/OTHER ။/ sb

ဦးသန့်/PER သည်/OTHER ၁၉၅၇/OTHER ခုနှစ်/OTHER မှ/OTHER ၁၉၆၁/OTHER ခုနှစ်/OTHER အထိ/OTHER မြန်မာနိုင်ငံ/LOC ၏/OTHER ကုလသမဂ္ဂ/ORG အမြဲတမ်း/OTHER ကိုယ်စားလှယ်/OTHER အဖြစ်/OTHER သာမက/OTHER အယ်လ်ဂျီးရီးယားနိုင်ငံ/LOC ၏/OTHER လွတ်လပ်ရေး/OTHER စေ့စပ်/OTHER ဆွေးနွေး/OTHER မှု/OTHER များ/OTHER တွင်/OTHER လည်း/OTHER ပါဝင်/OTHER ခဲ့/OTHER သည်/OTHER ။/sb

PER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER LOC OTHER OTHER OTHER OTHER

OTHER OTHER PER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER

PER OTHER LOC OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER LOC OTHER OTHER OTHER OTHER OTHER OTHER

OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER ORG OTHER OTHER OTHER OTHER

PER OTHER OTHER OTHER OTHER OTHER OTHER OTHER LOC OTHER ORG OTHER OTHER OTHER OTHER LOC OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER OTHER

Now we calculate all the parameters of HMM model. These are

States= { PER ,LOC, ORG, OTHER,}

Total Sentences = 5
Total words for PER = 4
Total words for LOC = 5
Total words for ORG = 2
Total words for OTHER= 77

TABLE.I START PROBABILITY(π)

PER	LOC	ORG	OTHER
3/5	0/5	0/5	2/5

TABLE.II TRANSACTION PROBABILITY(A)

	PER	LOC	ORG	OTHER
PER	0	0	0	4/4
LOC	0	0	0	5/5
ORG	0	0	0	2/2
OTHER	1/77	5/77	2/77	69/77

Emission Probability (B) =

Since in the emission probability we have to consider all the words in the file. But it's not possible to display all the words so we just gave the snapshot of first sentence of the file. Similarly we can find the emission probability of all the words.

PER = 1/4
LOC = 1/5
ORG = 0/2
OTHER = 13/77

V. PERFORMANCE EVALUATION

To evaluate the algorithm through accuracy, precision, recall and f-measure in table 2, there is a need to count true positives, false positive, true negative and false negatives in the result records [8] table1.

Precision: It is the fraction of the correct answers produced by the algorithm to the total answer produced. The formula for precision is:

Precision(P)
= (Corrected answers/ answers produced) (1.4)

Recall: It is the fraction of the documents that are matching to the query mentioned and are successfully retrieved. Recall is calculated in the following manner:

Recall (R) **=**
(Corrected answers/total possible answers) (1.5)

F-Measure: It is the harmonic mean of precision and recall. The F-Measure is calculated as:

F-Measure = (2* R* P)/ (R+P) (1.6)

Precision = TP/(TP+FP);

Recall = TP/(TP+FN);

F-Measure = 2*precision*recall/ (precision+ recall);

accuracy = (TP+TN)/ total-population(N)

TABLE.III CONFUSION MATRIX FOR A BINARY CLASSIFIER

N = 3150	Postive	Negative	
Training	TP = 2620	TN = 380	3000
Testing	FP = 19	FN = 131	150
	2639	511	

TP =True Postive
TN =True Negative
FP =False Postive
FN = False Negative

TABLE.IV MEASURE ON TEST DATA

Measures	Result
Accuracy	0.9523809523809523
Precision	0.9928003031451307
Recall	0.9523809523809523
F-Measure	0.9721706864564007

VI. CONCLUSION

Named Entity Recognition is a long-studied technology with a wide range of natural language applications NER systems have been developed for resource-rich languages like English with very high accuracies. But construction of an NER system for a resource-poor language like Myanmar language is very challenging due to unavailability of proper resources. Myanmar is no concept of capitalization which is the indicator of proper names for some other languages like English. If we perform Named Entity Recognition in HMM and also provide the ways to improve the accuracy and the performance metrics.

References

[1] A. Nayan, B. R. K. Raoand P. Singh, S. Sanyal, and R. Sanyal. Named entity recognition for indian languages. In IJCNLP, pages 97–104, 2008.

[2] S. K. Saha, S. Chatterji, S. Dandapat, S. Sarkar, and P. Mitra. A hybrid approach for named entity recognition in indian languages. In Proceedings of the IJCNLP08 Workshop on NER for South and South East Asian Languages, pages 17–24, 2008.

[3] A. Ekbal, R. Haque, A. Das, V. Poka, and S. Bandyopadhyay. Language independent named entity recognition in Indian languages. In IJCNLP, pages 33–40, 2008.

[4] Leah Larkey, Nasreen AbdulJaleel, and Margaret Connell, What's in a Name? Proper Names in Arabic Cross-Language Information Retrieval. CIIR Technical Report, IR-278,2003.

[5] B. Sasidhar#1, P. M. Yohan*2, Dr. A. Vinaya Babu3, Dr. A. Govardhan4" A Survey on Named Entity Recognition in Indian Languages with particular reference to Telugu" in IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011 available at : http://www.ijcsi.org/papers/IJCSI-8-2-438-443.pdf.

[6] Shilpi Srivastava, Mukund Sanglikar & D.C Kothari. "Named Entity Recognition System for Hindi Language: A Hybrid Approach" International Journal of Computational P(qi|qi−1P(qi|qi−1Linguistics(IJCL), Volume(2):Issue(1): 2011.Availableat:http://cscjournals.org/csc/manuscript/Journals/IJCL/volume2/Issue1/IJCL-19.pdf

[7] "Padmaja Sharma, Utpal Sharma, Jugal Kalita"Named Entity Recognition: A Survey for the Indian Languages"(Language in India www.languageinindia.com 11:5 May 2011 Special Volume: Problems of Parsing in Indian Languages.) Available at:

[8] Simple-guide-to-confusion-matrix-terminology.pdf https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/