AN AUTOMATED SYSTEM FOR TAMIL NAMED ENTITY RECOGNITION USING HYBRID APPROACH

Mrs.N.Jeyashenbagavalli
Student/CSE-PG
National Engineering College,
Kovilpatti,Tamilnadu,India.
jeyamca08@gmail.com

Dr.K.G.Srinivasagan,
Prof. & Head,/ CSE-PG,
National Engineering College,
Kovilpatti,Tamilnadu,India.
kgsnec@rediffmail.com

Mrs.S.Suganthi
Asst.Prof.,/ CSE-PG,
National Engineering College,
Kovilpatti,Tamilnadu,India.
krish.sugi1@gmail.com

***Abstract*- Named Entity Recognition is the process of identifying and recognizing named entities such as person, organization, location, date, time and money in the text documents. Named Entity Recognition is a subtask of Information Extraction. Information Extraction is the process of extracting the relevant data from documents. It is one of the research areas in Natural language processing. In this project implement a named entity recognizer using the hybrid approach that uses both Rule based and Hidden Markov Model in succession, which identifies only person, location and organization names respectively. Input data for proposed Named Entity Recognition system is any text document related to the any domain but limited size corpora respectively in Tamil language. In this system are tagging each word by using POS tagger and then imposing certain rules such as Lexical features and use some Gazetteers. HMM model using E-M algorithm is taken output data from trained as input to recognition system. The main purpose of this system identifies unknown entities and solves the problem of same name entity in different positions in the same document. The system is measuring the recall and precision parameters calculate the F-measure score. Goal of this project is to improve the performance of NER system to achieving high F-measure score.**

Keywords-NLP; NER; HMM;,POS tagging; Morphological analyzer .

I. INTRODUCTION

Natural Language Processing (NLP) is the one of the field of computer science and intelligence, the linguistics is concerned with human language and computer program interaction[1].NLP is the development of computational aspects of human language processing. It is the process of an Extraction of meaningful information from human spoken language input and producing local language output.NLP is an area of research that explores how computers can be used to understand and manipulate natural language text to do useful things. The higher level task in Natural Language Processing are Machine Translation(MT), Information Extraction(IE), Information Retrieval(IR), Automatic Text Summarization(ATS) and Question-Answering Systems(QA). Machine Translation is the major prompt research area across the country.

Information Extraction (IE) systems analyses unrestricted text in order to extract information about pre-specified types of events, entities or relationships[2]. This is the roots in the Natural Language Processing (NLP) area, the topic of structure extraction now engages many different communities across machine learning, information retrieval, database, web, and document analysis. Early extraction tasks are concentrated around the identification of named entities, like people and company names and relationship among them from natural language text.

Tami named entity recognizer would help in automatically identifying the unknown entities natural language understanding, machine translation, speech recognition, speech synthesis, part of speech tagging, and parsing applications[11]. The common man can also get in-depth information about the named entity in Tamil noun is identifying the proper noun of the word verbs from the software. The processes of named entity recognition obtain the identification and classification named entities in given sentence. The sentence occur in the ambiguity how to identify the word in the sentence and related to what kind of entities occurred in person versus common and person versus place name .The word correctly identify the proper noun and classify the correct categories in the entities.

In English identifying the organization of any other entity through the capitalization is possible. But Tamil there is lack of capitalization. Tamil word more affixes are added so the word does not easily identify the root word[3]. This is complex one in Tamil language. In the corpus find lot of ambiguity between common and proper nouns. For example the word such as s “தாமரை” belongs to different named entity categories flower and person name and this word such as “காசி” belongs to represent person name or place name. In this place named entity recognition is a difficult task[6]. In some case if the proper noun is consider sperately, it belongs

978-1-4799-3966-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICICA.2014.95

to one named entity type but if consider the same named entity it belongs to another named entity type. For instance “இந்திராகாந்தி பல்கலைக்கழகம்” is an organization name. Two entities present in the word. இந்திராகாந்தி is person name and பல்கலைக்கழகம் is an organization name.

II. RELATED WORK

Named Entity Recognition is the one of the tasks defined in MUC 6. There are many techniques followed in Named Entity tagging. Named Entity Recognition was done different language using kannada by Amarappa [1]. The techniques used include multi engine approach technique by Asif Ekbal [2] using Bengali language, They considered root of words, POS, combined word and POS , Dictionary of named entities as features to build the system. Pandianl[4] have built a Tamil NER system using contextual cues and E-M algorithm Named Entity system focus on so many challenge and how to solve the problem make a rules and CRF based on E-M algorithm done by malarkodi[6]. Named entity like Person name, Location name, Organization name and miscellaneous name for Indian languages are namely Bengali and Hindi using the Maximum Entropy (ME) framework done by Mohammad Hasanuzzaman [7]. The NER using Hidden Markov Model by Nusrat Jahan [8] an hybrid approaches such as rule based tagging for certain entities such as date, time, percentage and maximum entropy based approach for entities like location and organization (Raymond Chiong[9]) There was also a novel approach using concept based seeds (Shilpi Srivastava[10]) using maximum entropy markov model and Conditional Random Field(CRF) . Named Entity Recognition for Tamil Using Conditional Random Fields developed a model titled .Domain Focused NE Recognizer for Tourism Domain Conditional Random Fields (CRF) Approach on Tamil Language(Vijaya krishna R[11]). An Automated named entity recognizer system first describe the proposed work included problem statement .Then describe the frame work and various methodologies. Next make a rules in noun identification .

III. PROPOSED WORK

To develop NER system is to identify and classify the named entities. In Tamil language have agglutinative nature. So case markers attach as postpositions to proper or common nouns can form a single word. Ambiguities are occurred in the word. The special day name entity category consists of either month name or person name named entity embedded within it*(கிருஷ்ண ஜெயந்தி)*. This makes the name entity identification system to tag both separately instead of one single entity. The sentence represented in nested entities. Identification of named entities is important in several higher language technology systems such as information extraction systems, machine translation systems, and cross-lingual information access systems.

IV.NER SYSTEM FRAME WORK

Named Entity Recognition system frame work fig.1 includes several processing stages. The input sentence is a Tamil string. Figure 1 represents the frame work for a simple identify the unknown entities of the NER System[8]. NER System is processing a document using several of the procedures .First step, the raw text of the document is splitting into sentences using a sentence segmented, and each sentence is further subdivided into words using a tokenizer. The next stage is tokenized of each word is tagged using standardized parts of speech tag. Parts of speech tagging is one of the task in language processing. The next stage is tokenized of each word is tagged using standardized parts of speech tag. Parts of speech tagging is one of the task in language processing. It is the process of labeling the syntactic categories for each word in a corpus.

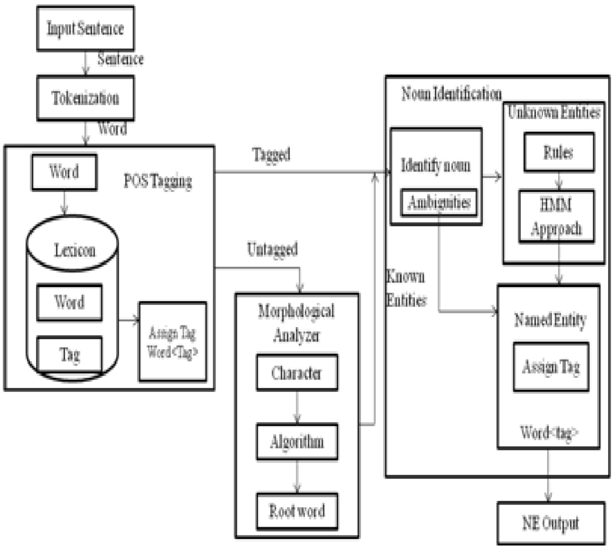


Fig 1. Named Entity Recognition Frame Work

Generally identification of ambiguities in language lexical items is the challenging objective in POS Tagging like noun, verb, adjective and adverb etc.The sequential labelling methods some words are not identified through the lexicon. The POS tag are used own tag set mention in Table 1

Table 1 POS Tag

S.No	NE Tag	Description
1	NN	Noun
2	V	Verb
3	ADJ	Adjective
4	ADV	Adverb
5	PRON	Pronoun
6	ORD	Ordinal

Example:

எழிலன்	கடைக்கு	சென்று	புத்தகம்	வாங்கினான்
NN	**Others**	**Others**	**NN**	**Others**

Then untagged words process into the morphological Analysis is the process of analyzing the internal pattern of a word and also it is very useful for identifying the root of a given word in a sentence[9]. The combinatorial technique underlies the morphological processing of word forms.

Example:

எழிலன் கடைக்கு சென்று புத்தகம் வாங்கினான்
NN NN V NN V

It analyses the naturally occurring word forms in a sentence and identifies the root word and its features. The general form morphological analyzer of Tamil word is root/stem and suffixes. The rule based forwarded algorithm is the way of removing the suffixes and extracting the morphemes of a word. The morpheme is the meaningful piece of a word. Noun identification is identifying the noun and assign the predefined categories of the tag. Otherwise name the category as "unknown word". There are various approaches used for entity identification such Rule based and Machine learning approaches. The following Table2 represent NER used in own Tag list.

Table 2 NER Tag list

S.No	NE Tag	Description
1	**NEP(Person Name)**	கிருஷ்ணன்,தேவி
2	**NEO((Organization)**	கூகிள் நிறுவனம், கார்ப்பரேஷன்
3	**NEL(Location)**	சென்னை,மதுரை
4	**NET(Time)**	வருடம்,10thஜூலை
5	**NEM(Measure)**	ஐந்து கிலோ,ஆறு கி.மீ
6	**NEN(Number)**	எழு,அறுபத்திமூன்று
7	**Others**	Unknown entities

Example

எழிலன் கடைக்கு சென்று புத்தகம் வாங்கினான்

NEP NEL Others Others Others

V. HYBRID APPROACH

Mainly there are two approaches for the named entity recognition task. Knowledge based approach is the process of defining set of rules by human experts. It takes lot of time and human effort because this process is done by the Human experts. Knowledge based approach is less efficient in recognizing the named entities. Another approach for entity recognition is Machine leaning approach [6]. Rule based approaches which lack robustness and portability then machine leaning approach used in Statistic Models for Hidden Markov model are very efficient in finding named entities.

The system using these two approaches in succession and achieve high F-measure score. Advantage of using these two models in succession is that, in HMM approach are mainly focus on ambiguous problem. Ambiguous refers to occurrence of the same entity name in many times within the document. The main purpose of using this model is to avoid the unnecessary counting of the same entity name in many times. Because, if the system count the same entity name many times then it leads to the poor performance of the system. Due to the system are achieving High F-measure score, hence NER system performance increases.

A. Rules-Based For Tamil NER

Named Entity recognition system overcome to the agglutination problem for considering and identifying the root word using orthographic rules and forwarded algorithm [6]. The rule helps to improve the identification of entities like person name but the entities are attached to several case marker in the root words for the system.

Rule 1

The end of the word is attached to the case marker in noun category. In Tamil word attached to suffixes are க்கு, உடைய, இல், ஐ, அது, ஆல், ஆகா, இன்.

Example:

மலர் கடைக்கு சென்று மலரை வாங்கினாள்

The two words to attain the agglunative problem and resolve the problem using orthographic rules.

மலரை= மலர் + ஐ

கடைக்கு=கடை+க்கு

Rule 2

To solve the problem in person name and location name. first word depends on the second word and the second words is location tag.

Example

முருகன் கோவில் கூட்டமாக உள்ளது

NEP NEL

முருகன் is the person name and கோவில் is the location name. The second word has another entity combine the two words and put the location tag. This rules used to nested entity. All the words in the sequence belong to a maximal NE tag and to assign the last NE tag in the sequence to the maximal NE.

Rule 3

To obtain the time and measure tag using rules. Time mention the word depends on previous word. The word present in நாள்,வருடம்,மாதம்..

Example:

கார்த்திகை மாதம் = NET

NEP NET

Rule 4

If a previous word of a particular token is a Pre-nominal word like (திரு.),(திருமதி),(செல்வி) etc then the word is a named entity.

Example: திரு.மதிவாணன்

If a word like (சாமி), (ராவ்) exist after a particular token then it denotes a person named entity.

Example: கந்தசாமி,ராமாராவ்

Rule 5

Words are continuously present the tag is name of the person.That word is name of the person entity tag.

Example:

தேவி ஸ்ரீ விஷ்ணு = **NEP**

NEP NEP NEP

Rule 6

The word end of the character with புரம், பட்டி represent the location name.

Example:

கோவில்பட்டி, விழுப்புரம் = **NEL**

The word end of the character is லிட், That word is organization name.

B. Hidden Markov Model (HMM)

Hidden Markov models (HMMs) are a powerful probabilistic tool for modeling sequential data, and have been applied with success to many text-related tasks, HMMs are probabilistic finite state models with parameters for state-transition probabilities and state-specific observation probabilities. In text-related tasks, the observation probabilities are typically represented as a multinomial distribution over a discrete, finite vocabulary of words, and E-M algorithm is used to learn parameters that maximize the probability of the observation sequences in the training data. There are two problems with this traditional approach. First, many tasks would benefit from a richer representation of observations in particular a representation that describes observations in terms of many overlapping features, such as word endings, part-of-speech, formatting, position on the page in Word[8].

The Expectation-Maximization (EM) iterative algorithm is a broadly applicable statistical technique for maximizing complex likelihoods and handling the incomplete data problem [6]. The EM algorithm works iteratively in alternatingly applying two steps: the E-Step (*expectation*) and the M-Step (*maximization*). Formally, $\theta^{(t)}$ for $t = 0; 1; 2; :::$, denote the successive parameter estimates; the E and M steps are defined as:

E-step:

Compute log p(y,**θ|x**) given the observed data x and the current parameter estimate $\theta^{(t)}$)

$$Q(\theta|\theta^{(t)}) \equiv E[log\,p(y,\theta|x)x,\theta^{(t)}]$$

$$\alpha\ logp((\theta)+E[logp(y,x|\theta)|x,\theta^{(t)}]$$

$$= logp(\theta)+\int p(y|x,\theta^{(t)})l\ ogp(y,x|\theta)dy.$$

M- step:

Select $\boldsymbol{\theta}^{\mathbf{k+1}}$ the value of θ that maximizes $\mathbf{Q(\theta \mid \theta^{k})}$ **is to** identify the correct category of the word in a sequence is the maximum probability of the word select from the next word to that category. The process continues until some stopping criterion is met.

$$\theta^{(t+1)} = \underset{\theta}{arg\,max}\ Q(\theta|\theta^{(t)})$$

VI. PERFORMANCE ANALYSIS

The performance of NER system can be computed using F-measure score. This based on the two parameter. They are Precision and Recall.

Recall:

Recall is defined as the number of correct tags in the document marked up by our proposed NER system over the total number of annotated tags present in the document.

It is denoted as "R" for our convenience. The main purpose of recall is to measure how well our system can perform the recognition of entity names.

Precision

Precision is defined as the number of correct tags in the file marked up by our system over the total number of tags being marked up. It is denoted as "P".

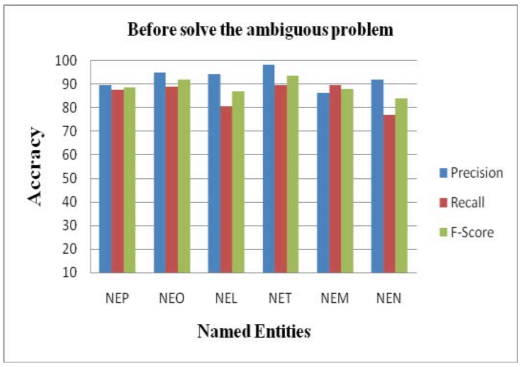
F-score= $\frac{2 * P * R}{P+R}$

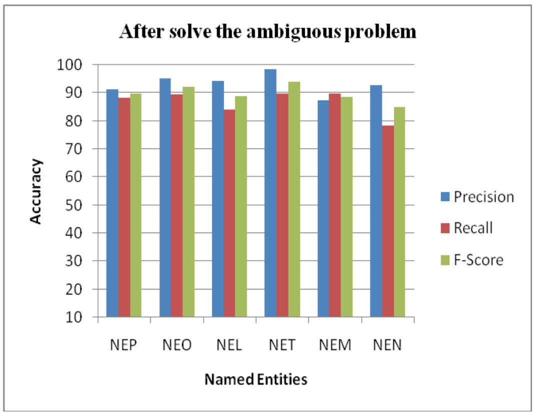
The corpus consists of 15,000 words. The corpus preprocessed for POS and Morphological analyzer is automatically analyze the tool. Named entities are identifies and results are evaluated manually, the following Table shows the evaluation results before solve the problem. Named entities solve the problem in after obtained the system is evaluated in Table3

Each named entities computed the precision and recall .These values are represents on the chart. These chart evaluated results obtained after the solve the problem using rules in named entity recognition system.

Table 3 After solve the ambiguous problem

NE Type	Precision(%)	Recall (%)	F-score (%)
NEP	91.2	88.3	89.7
NEO	95	89.4	92.1
NEL	94.26	83.9	88.8
NET	98.5	89.69	93.9
NEM	87.45	89.63	88.5
NEN	92.67	78.3	84.9
Over All	93.18	86.54	89.7





VII. CONCLUSION

The Named Entity Recognition (NER) system using hybrid approach that use both rule based and Hidden Markov Model in succession. The experimental results show that performance of NER system is better with Hybrid approach than using single statistic model for HMM. The named entity recognizer is most effective and efficiently to find the all the named entities in the sentence. The proposed system solves ambiguous, nested entities in the sentence. NER system has good performance in finding named entities and achieving high F-measure score with limited size corpora.

REFERENCES

[1] S Amarappa and S V Sathyanarayana ‘Named Entity Recognition and Classification in Kannada Language’ IJECSE Page No:281-289 March 2011.

[2] Asif Ekbal ‘ Named Entity Recognition in Bengali: A Multi-Engine Approach’ Northern European Journal of Language Technology, 2009, Vol. 1, Article 2, pp 26–58.

[3] Deepti Chopra, Nusrat Jahan, Sudha Morwal, ’ Named Entity Recognition by Aggregating Rule Based Heuristics And Hidden Markov Model Hindi’, IJIST Vol.2, No.6, November 2012.

[4] Lakshmana Pandian,Krishnan Aravind Pavithra,T.V. Geetha ‘Hybrid, Three-stage Named Entity Recognizer for Tamil’ INFOS2008, March 27-29, 2008 Cairo-Egypt © 2008 Faculty of Computers & Information-Cairo University.

[5] Lakshmana Pandian S and Kumanan (2012), ‘Machine translation from English to Tamil using Hybrid Technique’ IJCA, Vol.46, .16 ,Pages: 36-42.

[6] Malarkodi, C S., Pattabhi, RK Rao and Sobha, Lalitha Devi ‘ Tamil NER – Coping with Real Time Challenges’ Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 23–38.

[7] Mohammad Hasanuzzaman, Asif Ekbal and Bandyopadhyay ‘Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi’ International Journal of Recent Trends in Engineering, Vol. 1,No.1, May 2009.

[8] Nusrat Jahan , Sudha Morwal and Deepti Chopra ‘Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model: A Hybrid Approach’ IJCSET , Vol. 3 No. 12 Dec 2012 pp no:621

[9] Raymond Chiong and Wang Wei ,’Named Entity Recognition Using Hybrid Machine Learning Approach’ December 11, 2008 at 06:15 from IEEE Xplore.

[[10] Shilpi Srivastava, Mukund Sanglikar and D.C Kothari, ‘Named Entity Recognition System for Hindi Language: A Hybrid Approach’ IJCL, Volume (2) : Issue (1) : 2011

[11] Vijayakrishna R, Sobha L ‘Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields’ IJCNLP-08,pages 59–66, January 2008.

[12] Vishal Gupta, and Gurpreet Singh Lehal, ‘Named Entity Recognition for Punjabi Language Text Summarization’ International Journal of Computer Applications (0975 – 8887) Volume 33– No.3, November 2011.