Named Entity Recognition for Telugu Using Conditional Random Field

G.V.S.Raju letter2raju@gmail.com
Professor of CSE Department
Indur institute of Engg.&Tech
Siddipet,A.P, India

B.Srinivasu srinivas_534@yahoo.com
Asso Professor of CSE Department
Indur institute of Engg.&Tech
Siddipet,A.P, India

S.VISWANADHA RAJU viswanadharajugriet@gmail.com
Professor of CSE Department
JNTUH College of Engineering
Jagityala , A.P, India

ALLAM BALARAM balaramallam@gmail.com
Asst professor of CSE Department
Indur institute of Engg.&Tech
Siddipet,A.P, India

Abstract

Named Entity (NE) recognition is a task in which proper nouns and numerical information are extracted from documents and are classified into predefined categories such as Person names, Organization names , Location names, miscellaneous(Date and others). It is a key technology of Information Extraction, Question Answering system, Machine Translations, Information Retrial etc. This paper reports about the development of a NER system for Telugu using Conditional Random field (CRF). Though this state of the art machine learning technique has been widely applied to NER in several well-studied languages, the use of this technique to Telugu languages is very new. The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the four different named entities (NE) classes, such as Person name, Location name, Organization name, miscellaneous (Date and others).

Keywords: Named entity, Conditional Random field, NE, CRF, NER, named entity recognition,Telugu

1. INTRODUCTION

Named Entity (NE) recognition is an important tool in almost all natural language processing applications like Information Extraction (IE), Information retrieval and machine translation and Question answering system etc. The objective of NER is detect and classify each and every word or token in a text document into some predefined categories such as person name, location name, organization name, date and designation. Identification of named entity is a difficult task because named entities are open class expressions, i.e there is an infinite verities and new expressions are constantly being invited.

NER system has been developed for resources rich language like English is very high accuracies. But development of NER system for a resource poor language like Telugu is very challenging due to unavailability of proper resources.[5]English is resource-rich language containing lots of resources for NER and other NLP tasks. Some of the resources of English language can be used to develop NER system for a resource-poor language. Also English is used widely in many countries in the world. In India, although there are several regional languages like Telugu, Kannada, Tamil, Hindi etc.., English is widely used (also as subsidiary official language). Use of the Telugu languages in the web is very little compared to English and other Indian languages. So, there are a lot of resources on the web, which are helpful in Telugu language NLP tasks, but they are available in English. For example, we found several relevant name lists on the web which are useful in Telugu NER task, but these are in English. It is possible to use these English resources if a good transliteration system is available.

Transliteration is the practice of transcribing a word or text in one writing system into another. Technically most transliterations map the letters of the source script to letters pronounced similarly in the goal script. Direct transliteration from English to an Telugu language is a difficult task.

A large number of techniques have been developed to recognize named entities for different languages. Some of them are Rule based and others are Statistical techniques. The rule based approach uses the morphological and contextual evidence (Kim and Woodland,2000) of a natural language and consequently determines the named entities. This eventually leads to formation of some language specific rules for identifying named entities. The statistical techniques use large annotated data to train a model (Malouf, 2002) (like Hidden Markov Model) and subsequently examine it with the test data. Both the methods mentioned above require the efforts of a language expert. An appropriately large set of annotated data is yet to be made available for the Indian Languages. Consequently, the application of the statistical technique for Indian Languages is not very feasible. This paper deals with a CRF technique to recognize named entities of Telugu languages.

2. NER FOR INDIAN LANGUAGES

NLP research around the world has taken giant leaps in the last decade with the advent of effective machine learning algorithms and the creation of large annotated corpora for various languages. However, annotated corpora and other lexical resources have started appearing only very recently in India. Not much work has been done in NER in Indian languages in general and Telugu in particular. Here we include a brief survey.

In (Eqbal, 2006), a supervised learning system based on pattern directed shallow parsing has been used to identify the named entities in a Bengali corpus. Here the training corpus is initially tagged against different seed data sets and a lexical contextual pattern is generated for each tag. The entire training corpus is shallow parsed to identify the occurrence of these initial seed patterns. In a position where the seed pattern matches wholly or in part, the system predicts the boundary of a named entity and further patterns are generated through bootstrapping. Patterns that occur in the entire training corpus above a certain threshold frequency are considered as the final set of patterns learned from the training corpus.

In (Li and McCallum, 2003), the authors have used conditional random fields with feature induction to the Hindi NER task. The authors have identified those feature conjunctions that will significantly improve the performance. Features considered here include word features, character n-grams (n = 2,3,4), word prefix and suffix (length - 2,3,4) and 24 gazetteers.

3. SOME CASES IN TELUGU LANGUAGE NAMES

Some of the typical ambiguous cases in Telugu
variation of Named Entities:
వైయస్ రాజశేఖర్ రెడ్డి వైయస్, వై.యస్.ర్
vaiyas raajas`eekhar reDDi, vaiyas, vai.yas.r.
Y.S. Rajashakar Reddy, Y.S. Y.S.R

Ambiguity in NE type:

సత్యం (SatyaM) person Vs organizatio
తిరుపతి (Tirupati) person Vs location

Ambiguity with Common Noun:

బంగారు Gold (baMgaaru) person first Vs common noun

Appearance in various forms:

తెలుగుదేశంపార్టీ, టీడీపీ, తె.దే.పా
telugudees`aMpaarTii, TiiDiiPii, te.dee.paa
Telugu Desam party, T.D.P, Te.De.Pa

These are some examples which show the complexity of development of NER system .This paper is focused on development of NER for Telugu Language

4. APPROACHES FOR NER

Named entity recognition is a classification and identification problem but this is a kind of problem which requires features of the word at least in case of Telugu for proper identification of NEs. Widely used approaches for solving such problems are Statistical Machine learning Techniques, Rule Based System or hybrid approach. In Statistical techniques many approaches may be applied like Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM) and Conditional Random Field (CRF) [3], support vector machine(SVM). Sequence labeling problem can be solved very efficiently with the help of Markov Models. The conditional probabilistic characteristic of CRF and MEMM are very useful for development of NER system. Between both, MEMM is having bias labeling problem which makes it vulnerable for this research. CRF is flexible enough to capture many correlated features, including overlapping and non Independent features. Thus multiple features can be used in CRF more easily than in HMM.

All Machine Learning Techniques require a large relevant corpus which can pose a problem in case of Telugu and other Indian Languages because of unavailability of such corpus. The positive side of these techniques is being cost effective and requires less language expertise whereas Rule based system requires language expertise for crafting rules. A rule based system incurs huge cost and time.

5. DEVELOPMENT OF TAGGED DATA

News articles generally start with a headline and the body starts with location name, month and date. A seed list of location names is extracted from this. Seed lists of personal surnames, location names and organization names have also been developed. Lists of person suffixes such as “reDDi(reddy)”, “naayuDu(nayudu)” etc, location suffixes such as “baad”, “peeTa” “paTnaM” etc, and name context lists are maintained for tagging the corpus. It has been observed that whenever a context word (such as ”maMtri”) appears, then in many cases the following two words (consisting of a surname and person name) indicate a person name. This way we build a list of person names. After extensive experimentation over many iterations, a training data set of 30,000 words has been developed.

6. NOUN IDENTIFICATION

It is useful to recognize nouns and eliminate non-nouns. The Telugu morphological analyzer developed here has been used to obtain the categories. A stop word list including function words has been collected from existing dictionaries and stop words are removed. Words with less than three characters are unlikely to be nouns and so eliminated. Last word of a sentence is usually a verb(Telugu is verb final language, in ever sentence final word may be a verb) and is also eliminated. Digits are eliminated. Verbs are recognized based on a list of verb suffixes and eliminated. Telugu words normally end with a vowel and consonant ending words (laMDan(Landan), sTeeSan(station), meenejar (manager) etc.) are usually nouns. Existing dictionaries are also checked for the category. Using these features, a naive Bayes classifier is built using the available tool WEKA. Results are given in the tables 1 and 2.

	noun	not-noun
Precision	94.56	63.47
Recall	62.78	94.45
F-measure	76.25	75.38

TABLE 1: Noun Identification using Morphological Analyzer

	Test set-1		Test set-2	
	noun	not-noun	noun	not-noun
Precision	91.56	95.56	76.47	89.2
Recall	96.15	91.45	90.45	74.48
F-measure	95.1	93.67	83.28	81.15

TABLE 2:. Noun Identification using a naive Bayes Classifier

7. PROPOSED METHODOLOGY FOR NER

7.1 Labeling Sequential Data

The task of assigning label sequences to set of observation sequences arises in many fields, including speech recognition, computational linguistics. For example, consider the natural language processing task of labeling the words in a sentence with their corresponding part-of-speech (POS) tags. In this task, each word is labeled with the tag indicating it's appropriate part-of-speech resulting in annotated text such as :

<NPER>caMdrabaabu naayuDu <NLOC>raajoli graamamunu <V> saMdars`iMcaaru.

Chandrababu Naidu visited Rajoli village .

Labeling sentences in this way is a useful prepossessing step for higher NLP tasks: POS tags augment the information contained with in the words alone by explicitly indicating some of the structure inherent in the language.

One of the most common methods for performing such labeling and segmenting tasks is that of employing hidden Markov models (HMMs) or probabilistic finite-state automata to identify the most likely sequence of labels for the words in a given sentence. HMMs are a form of generative model, that defines a joint probability distribution p(X|Y) where X and Y are random variables respectively ranging over observation sequences and their corresponding label sequences. In order to define a joint distribution of this nature, generative models must enumerate all possible observation sequences a task which, for most domains,is intractable unless observation elements are represented as isolated units,independent from the other elements in an observation sequence. More precisely, the observation element at any given instant in time may only directly depend on the state, or label, at that time. This is an appropriate assumption for a few simple data sets, however most real-world observation sequences are best represented in terms of multiple interacting features and long-range dependencies between observation elements.

This representation issue is one of the most fundamental problems when labeling sequential data. Clearly, a model that supports tractable inference is necessary, however a model that represents the data without making unwarranted independence assumptions is also desirable. One way of satisfying both these criteria is to use a model that defines a conditional probability p(Y|X) over label sequences given a particular observation
sequence x, rather than a joint distribution over both label
and observation sequences. Conditional models are used to label a novel observation sequence x_* by selecting the label sequence y_* that maximizes the conditional probability p(y_*| x_*) The conditional nature of such models means that no effort is wasted on modeling the observations, and one is free from having to make unwarranted independence assumptions about these sequences; arbitrary attributes of the observation data may be captured by the model, without the modeler having to worry about how these attributes are related. Conditional random fields [lafferty](CRFs)are a probabilistic framework for labeling and segmenting sequential data, based on the conditional approach described in the previous paragraph. A CRF is a form of undirected graphical model that defines a single log-linear
distribution over label sequences given a particular observation sequence. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference.[D. Pinto,F. Sha, lafferty]

7.2 Undirected Graphical Models

A conditional random field may be viewed as an undirected graphical model, or Markov random field, globally conditioned on X, the random variable representing the observation sequences. Formally, we define G = (V,E) to be an undirected graph such that there is a node v in V corresponding to each of the random variables representing an element Yv of {Y}. If each random variable Y{v} obeys the Markov property with respect to G, then (Y,X) is a conditional random field In theory the structure of graph G may be arbitrary, provided it represents the conditional independencies in the label sequences being modeled. However, when modeling sequences, the simplest and most common graph structure encountered is that in which the nodes corresponding to elements of Y form a simple first order chain as illustrated in figure 1.

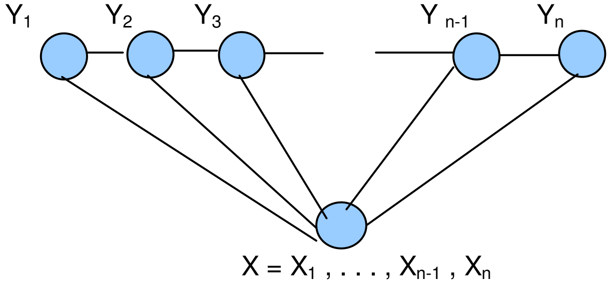

FIGURE 1

Let X is a random variable over data sequences to be labeled and Y is a random variable over corresponding label sequences. All components Y_i of Y are assumed to range over a finite label alphabet Y. For example, X might range over natural language sentences and Y may range over part-of-speech tagging of those sentences, with Y being set of possible part-of-speech tags.

Definition : Let G = (V,E) be a graph such that Y= Y_v, vεV, so that Y is indexed by vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X the random variables $Y_{v,}$ obey the Markov property with respect to graph : p(Y_v |X, Y_w, w≠ v)= p(Yv| X, Yw, w • v) where w • v means w and v are neighbors in G.
A CRF is a random field globally conditioned on the
observation X.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{\Psi_A \in G} \exp\left\{ \sum_{k=1}^{K(A)} \lambda_{Ak} f_{Ak}(\mathbf{y}_A, \mathbf{x}_A) \right\}$$

7.3 Suffix List:

a suffix list of Indian surnames inspire of using a specific surname. Similarly we can garner more and more features according to attributes of a language. Every feature function fi in CRF is having any real value on the basis of observation of the given language and these characteristics functions hold true for whole model distribution too.

Features of Telugu Language can be exploited for development of a good Named Entity Recognizer. Some features considered are as:

1. Some specific suffixes
2. Context Feature
3. Context Word List
4. Part of Speech of Words etc.

CRF has capability to introduce these features as binary value which makes it more useful for such problem.

7.3.1 **Features:**

1. Context Features: Preceding and following words of the current target word is very helpful for identifying NE category of the word. With identification of optimal window size of tokens we can get good results. For our experiment we have taken window size of five[2].
2. Context Pattern: Every language uses some specific patterns which may act as clue words and the list of this type of words is called as Context Lists. Such a list is compiled after analyzing Telugu text e.g. maMtri naayuDu,reDDi,adhyakshuDu etc for identification of person names and similarly for identification of places jilla, graamamu, bad, nagaraM etc.
3. Part-of-speech Features: Named Entities will fall in Noun Phrases and these boundaries can be found with help of Part of Speech category. Usually Verbs and Post-Positions denote the boundaries of such chunks. Some set of tags will give clue of being a word as NE.

External Resources (GAZ) : In order to measure the impact of using external resources in the NER task we have used A NERgazet which consists of three different gazetteers, all built manually using web resources:

(i) Location Gazetteer: this gazetteer consists of 22,000 names of villages, mandalas, Dicts, cities, in Andhra Pradesh found in the Telugu wikipedia and cities and states and countries found in other websites

(ii) Person Gazetteer: this was originally a list of 2000 complete names of people found in wikipedia and other websites. After splitting the names into first names and last names and omitting the repeated names, the list contains finally 3,450 names;

(iii) Organizations Gazetteer: the last gazetteer consists of a list of 400 names of companies,cricket teams,political party named and other organizations.

8. RESULT

We conducted experiments on a testing data of 150 sentences whereas model was created on 3000 sentences. Results on various combination's are tabulated in table 3, table 4, and table 5. Notations used in tables are as:

cw ->current word , pw ->previous word , nw->next word pw2->previous to previousword ,nw2 ->next to next word pt -> NE tag for previous word ,pt1-> NE tag for previous
to previous word cp –> Current pos tag , np -> next pos tag pp-> previouspos tag

Feature	Person	Location	Organization
pw , cw ,nw	57.8%	63.7%	40%
cw,pw,pw2,nw, nw2	60.4%	68.7%	48.7%
cw,pw,pw2,pw3 , nw, nw2, nw3	56.4%	67%	42.5%

TABLE3: Surrounding and current words combination:

Above table shows that window size of five gives optimum NE recognition. There would be no improvement in the accuracy even if the window size is increased further.

As shown in table4 Results are improved as compared to previous case because in this case we included NE tags which disambiguate some confusing classifications like in case of organization names

Feature	Person	Location	Organization
cw, pw, nw, pt	60.2%	64.6%	52.3%
cw, pw, pw2, nw, nw2, pt, pt2	62.2%	71%	52.00%
cw,pw, pw2,pw3, nw, nw2,nw3, pt, pt2,pt3	61.2%	69%	46%

TABLE 4: After adding NE tags:

Features	Person	Location	Organization
cw, pw, pp, nw, np, pt	66.7%	69.5%	58%
cw, pw, pp, pw2, pp2	66.3%	68%	58%

TABLE 5: After adding POS Tags:

Above table describes if we add pos tags in our word window, results could be improved further.

Above all results show that accuracy in case of organization, whatever combination we have taken ,is quite low compared to other NEs. It is because in most of cases organizations are multi word and even some cases comprise of Person Name, Location Name too. After inducting Part of Speech tag in feature, results get improved which justifies the approach chosen for embedding the features of language.

9. CONCLUSION

It is observed from our experiments and works done by other researchers too that CRF based system can be a viable solution if we identify and exploit features available in languages properly. It is also evident from experimental results (Table 5) that POS tags are playing an important role for getting better result. Our result is base on our training data and testing data. Most of the Indian languages gold standard data is not available because of unstable transliteration methods are used develop the training and testing data.

Future works includes increasing the relevant corpus size, induction of some more classification tags, and identification of some more features for Hindi Language. Inconsistency in the writing style e.g. telugu deesaM(Telugu Desam) and telugudeesaM (TeluguDesam) is the current limitation of the system, which can be handled too some extent by incorporating spelling normalization. Better classification of NEs can be achieved by induction of nested tag set. Rules can be crafted for identification and classification for the time, date, percentage, currency etc.

10. REFERENCES

[1] Asif Ekbal et. al*. "Language Independent Named Entity Recognition in Indian Languages".* IJCNLP, 2008.

[2] Prasad Pingli et al. *"A Hybrid Approach for Named Entity Recognition in Indian Languages".* IJCNLP, 2008.

[3] Lafferty, McCallum, et al. *"Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data".* 2001 .

[4] Himanshu Agrawal et. al*. "Part of Speech Tagging and Chunking with Conditional Random Fields".* IJCNLP, 2008

[5]. Lafferty J., McCallum A., and Pereira F. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In Proceedings of the Eighteenth International Conference on Machine Learning. 2001.

[6]. CRF++: Yet Another CRF toolkit http://crfpp.sourceforge.net/ (accessed on 13 rd Feb 2009)

[7] http://en.wikipedia.org/wiki/Named_entity (accessed on 11th Feb 2009)

[8]Navbharat Times http://navbharattimes.indiatimes.com (accessed on 11th Feb 2009)

[9] Chinchor, N. 1997. MUC-7 *Named entity task definition*. In Proceedings of the 7th Message Understanding Conference (MUC-7)

[10] Finkel, Jenny Rose, Grenager, Trond and Manning, Christopher. 2005. *"Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling."* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

[11] Kim, J. and Woodland, P.C. (2000a) *"Rule Based Named Entity Recognition".* Technical Report CUED/F-INFENG/TR.385, Cambridge University Engineering Department, 2000.

[12] Malouf, Robert.2002 *Markov models for language-independent named entity recognition*. In Proceedings of CoNLL-2002 Taipei, Taiwan, pages 591-599.

[13] Pramod Kumar Gupta, Sunita Arora, *An Approach for Named Entity Recognition System for Hindi: An Experim-ental Study*, Proceedings of ASCNT – 2009, CDAC, Noida, India, pp. 103 – 108

[14] T. W. Anderson and S. Scolve, *Introduction to the Statistical Analysis of Data*. Houghton Mifflin, 1978.

[15] Kristjansson T., Culotta A., Viola P., and McCallum A. 2004. *Interactive Information Extraction with Constrained ConditionalRandom Fields.* In Proceedings of AAAI-2004.

[16] D. Roth and W. Yih. *Integer linear programming inference for conditional random fields*. In Proc. of the International Conference on Machine Learning (ICML), pages 737–744, 2005

[17] Zobel, Justin and Dart, Philip. 1996. *Phonetic string matching: Lessons from information retrieval.* In Proceedings of the Eighteenth ACM SIGIR International Conference on Research and Development in Information Retrieval, Zurich, Switzerland, August 1996, pp. 166-173.

[18]. Li W. and McCallum A. 2003. *Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. In Special issue of ACM Transactions on Asian Language Information Processing*: Rapid Development of Language Capabilities: The Surprise Languages.

[19]. F. Sha and F. Pereira. *Shallow parsing with conditional random fields. roceedings of Human Language Technology,* NAACL 2003, 2003.

[20]. D. Pinto, A. McCallum, X. Wei, and W. B. Croft. *Table extraction using conditional random fields.* Proceedings of the ACM SIGIR, 2003.

[21]. Charles Sutton,Andrew McCallum, *An Introduction to Conditional Random Fields for Relational Learning*, Department of Computer Science University of Massachusetts, USA

[22]. Paul Viola and Mukund Narasimhan. *Learning to extract information from semistructured text using a discriminative context free grammar*. In Proceedings ofthe ACM SIGIR, 2005.

[23]. G.V.S.Raju, B.Srinivasu, S.V.Raju and Kumar*, Named Entity Recognition For Telugu using maximum entropy Model* , Journal of Theoretical and Applied Information Technology (JATIT), Vol-13, No-2, pages 125-130.