NAMED ENTITY IDENTIFICATION AND CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Dr.M.Humera Khanam[1], Mr.Md.A.Khudhus[2]
[1]Dept of Computer Science and Engineering, Sri Venkateswara University, Tirupati
[2]BSNL, Tirupati,

ABSTRACT-*In this paper, we discuss Named Entity Identification and Classification Using Machine Learning Techniques for Telugu Language. Identifying Named Entities is also known as Named Entity Recognition (NER).Named Entity Recognition is a task in which entities like proper nouns and numerical information is extracted from documents and classified into predefined categories such as person names, organization names, place, date, time, miscellaneous names. Here, in this paper we are using hybrid approach i.e. combination of rule based approach and one of the machine learning techniques (conditional random fields algorithm) to develop named entity recognizer. The identification and classification of entities often involve ambiguities. In order to resolve the ambiguities we have to choose the most appropriate tag from the valid tags available for the entities. In our system we are trying to improve the accuracy using CRF's. If we use solely a rule-based approach we can process very fast using pre-defined rules but ambiguity cannot be resolved and if we use solely machine learning technique it can process using annotated training data but maintaining training data is difficult. So in our proposed system we are using a hybrid approach to develop the accuracy of the system. Initially input is given in the form of paragraphs which are later converted into sentences using rule-based approach and later tokenized using a tokenizer, then go for direct matching. After matching, tag name is given to the particular word if found either in the dictionary or in gazetteer lists otherwise, it is declared as unknown word. Previously, the accuracy in the existing system was 84%.In our system, we are getting the accuracy about 90%.*

Keywords - Information Retrieval (IR), Conditional Random Fields (CRFs), Natural Language Processing (NLP), Named Entities (NEs).

1 INTRODUCTION

Natural Language Processing (NLP) is the use and ability to process the sentences in natural languages such as Telugu, Tamil, English, Hindi etc., rather than in specialized computer Languages like c, c++, java etc. NLP is concerned with a language of document for providing a semantic view. The ultimate goal of NLP is to provide Human Computer Interaction (HCI). Languages are dynamic in nature and contain Inclusions and deletions with the developments of sets of well-defined rules. Development of Named Entity Recognizer (NER) is very difficult for Telugu because of lack of resources such as gazetteers list, dictionaries and also different accent of people belonging to different regions. In languages like English, capitalization is a big resource for identification of proper nouns. The feature of capitalization is not available in Indian languages, particularly, in Telugu. There are very few resources openly available for carrying out a research work on Named Entity Recognition. Due to the above reasons, till now the accuracy for the NER system in Telugu is 85%(approx) whereas the accuracy in Hindi had reached to 99%.So we are trying to improve the accuracy to more than 90% .

2. RELATED WORK

The history of NLP started in the year 1950, although work can be found from earlier periods. In 1950, Alan Turing published his famous article "Computing Machinery and Intelligence", he proposed now popularly called the Turing test as a criterion of intelligence. Some successful NLP systems developed in the 1960s were SHRDLU ,a natural language system working in restricted "blocks worlds" with restricted vocabularies ,and ELIZA, a simulation of a Rogerian psychotherapist, written by Joseph Weizenbaum around 1964 to 1966.In 1970 's many programmers written conceptual ontologies which structured real-world information into computer-understandable data. Up to the 1980's most NLP systems were based on complex sets of hand-written rules. Many of the notable early successes occurred in the field of machine translation, due especially to work at IBM Research, where successively more complicated statistical models were developed. Recent research has increasingly focused on unsupervised and semi-supervised learning algorithms. In recent studies, techniques like maximum entropy model, Hidden Markov Models were used. Until now the accuracy obtained for Telugu language is just 80% but for Hindi and Bengali it is nearly 98%.Our proposed system consists of CRF tool to resolve ambiguity. Using CRF solely we can resolve most of the ambiguities. So we are using a hybrid approach which is a combination of rule-based approach and machine learning approach (CRF’s).

3. APPROACHES

Approaches for developing NER are categorized into two categories. They are described below in detail.

1. Rule based Approach.
2. Using Machine Learning Technique

3.1. Rule based Approach

Rule based Approach is based on rules which are written manually. Rules are written in the form of gazetteer lists which requires large amount of grammatical knowledge and experience regarding the language. This approach is language specific i.e. for different languages we have to maintain different set of rules.

Advantage: This approach is simple and fast.

Disadvantage: It is difficult to maintain gazetteer lists and it cannot resolve ambiguities.

3.2. Using Machine Learning Technique

Machine Learning techniques are based on annotated data to train the model. Some of the algorithms which comes under Machine Learning Techniques are:

- Hidden Markov Models (HMM)
- Maximum Entropy Markov Models (MEMM)
- Conditional Random Fields (CRF)
- Support Vector Machines (SVM)
- Decision Trees (DT)

3.2.1 Hidden Markov Models (Hmm)

It is a generative model. It assigns joint probability to paired observation and label sequence. It is advantageous because its basic theory is elegant and easy to understand. In order to define joint probability over observation and label sequences HMM needs to enumerate all observation sequences.

3.2.2 Maximum Entropy Markov Model (MEMM)

It is a statistical model which assigns an outcome for each token based on its history and features. The probability of each class is obtained for each word. To find most probable tag corresponding to each word of a sequence we can choose the tag having highest class conditional probability value.

3.2.3 Conditional Randomm Fields (CRFs)

It is a discriminative probabilistic model. CRFs are undirected graphical models which are used to calculate conditional probability of value on assigned output nodes given the values assigned to other input nodes.

3.2.4 Support Vector Machine (SVM)

The goal of SVM classifier method is to produce a model which predicts target value of the attributes. SVM maintains two data sets namely training and testing where SVM use the training set to make a classifier model and classify testing data set based on this model with use of their features.

3.2.5 Decision Tree

Decision Tree is a classifier in the form of a tree structure where each node is either a leaf node which indicates the value of the target attributes of expressions, or a decision node that specifies some text to be carried out on a single attribute value with one branch and sub-tree for each possible outcome. Advantage is no need to maintain gazetteers list and predefined rules. Disadvantage: Maintaining training data is bit difficult and ambiguities are not resolved completely.

4. PROPOSED SYSTEM

In our proposed system we are using hybrid approach in which we use a machine learning technique along with rule-based approach.

4.1 Hybrid Approach

Hybrid approach is a combination of rule-based approach and machine learning technique. We use this hybrid approach in order to improve accuracy because it is not easy to resolve ambiguity by using either rule-based approach or machine learning technique solely. Advantage is improved accuracy.

The process ofNER system is described in three phases.

a) Noun Identification and

b) Named Entity Identification and classification. c) Stemming process, where we use CRF's tool to remove ambiguity.

Noun Identification is done using rule based approach whereas Named Entity Identification and classification is done using machine learning technique i.e. conditional random field.

4.2 Noun Identification Using Rule Based Approach

Noun Identification is a easy process for English language when compared to non-phonetic languages such as Telugu, Urdu etc., because of availability of rich resources like dictionaries, gazetteer lists ,capitalization feature etc., For these non-phonetic languages we have to create these resources gazetteer lists Transliteration technique is very helpful for the preparation of gazetteers lists in Indian languages. We have developed a transliteration based gazetteers collected from various sources. This approach requires the loading of the computer system with a dictionary of Telugu language .It is helpful for identification of closed class words, such as, adjectives, adverbs, verbs, conjunctions and root forms nouns. Nouns may also appear with various inflected forms. We have to identify the nouns by using various suffix features and rules.

4.2.1 Steps for Noun Identification

1. Reading each file (document) and dividing it into sentences (pre-processing and cleaning the input file).
2. Divide each sentence and split it into tokens.
3. If the word is found in a Telugu dictionary, then assign it a category name.

4. The Last word of a sentence is usually a verb (Telugu is verb-final language in general. It is generally observed that most of the sentences end with 90% of verbs and in every sentence the final word may be a verb but not a noun.
5. If the word is not matched with dictionary, check the suffix match with the noun suffix list. If the word is found in the noun suffix list, then stemming process is applied to remove the suffix and again search the dictionary
6. If the word is not found again in the dictionary then the word may be a noun or guessing noun.
7. All Telugu words normally end with a vowel and consonant ending words are usually guessing nouns or loan words.
8. Otherwise name the category as “unknown word”.

Some of the common noun suffixes are:

Words in telugu	Transliterated form	Meaning
కి	ki	to
కు	ku	to
తో	tho	with
నుండి	nundi	from
తోపాటు	thopaatu	along with
లో	lo	from
చే	chey	with
ని/లని	ni/lani	in
ను/లను	nu	for
లతో	lato	with
లచే	lache	along
లకు	laku	for
గుండా	gunda	from

In the first phase we have conducted various tests for noun identification using dictionary gazetteers lists, suffix mapping techniques and other features. The results are highly appreciative. A great percentage of nouns typically of the order of 90% are identified in the first phase. These words which are not identified may be Named Entities or loan words. The identified nouns are given as input to the second phase to identify and classify the NE's.

4.3 Named Entity Identification and Classification

4.3.1 Steps for Named Entity Identification and Classification

1. Read input (only guessing nouns and unknown words) from Noun identification.
2. Check each word in the NE lists (Person, Location, and Organization).
3. If the word is found in the list then assign appropriate tag.

4. If not found, then check NE suffix list.
5. If the word is matching with any one of the suffix, then assign the appropriate tag.
6. Else, assign the category as “Un known word”.
7. If there is any ambiguity with more than one tag, then call disambiguation technique and remove ambiguity using some
Pattern matching techniques, suffix, context lists and assign the category, otherwise “unknown word”.

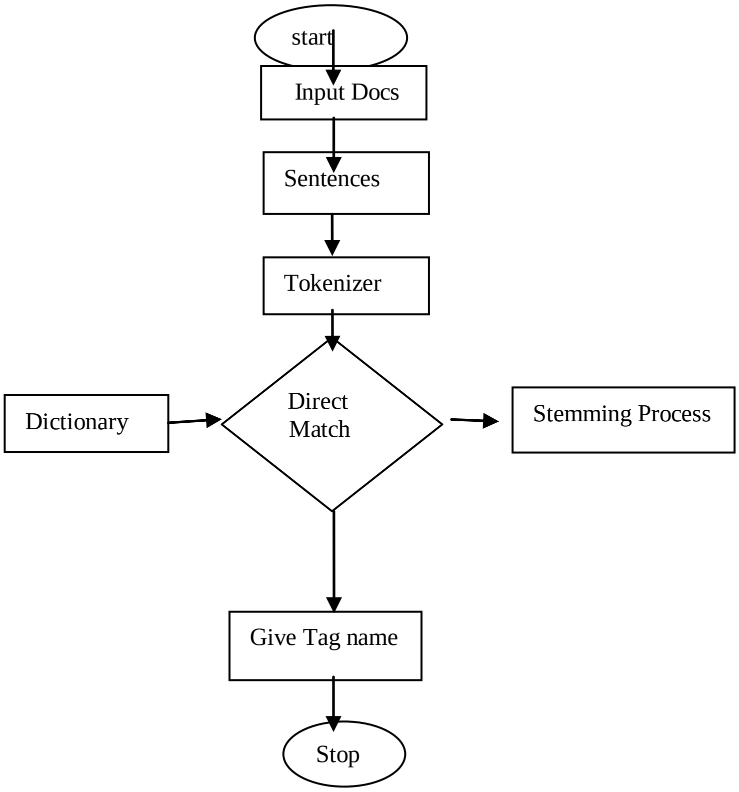


Fig 4.3.1 Flow Chart for Noun Identification

In the second phase we checked each noun with gazetteers lists which contain beginnings, endings, contexts and suffixes of various tags. According to the category NE tags are assigned in each category. Ambiguities are also resolved by using gazetteers lists and features. After conducting NE identification it is observed that great results are achieved by the system.

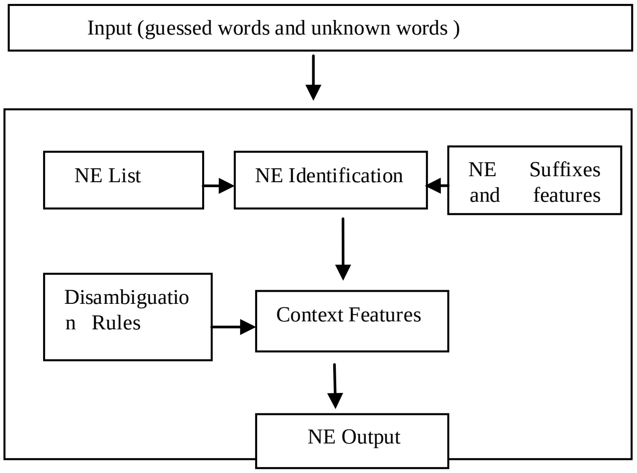
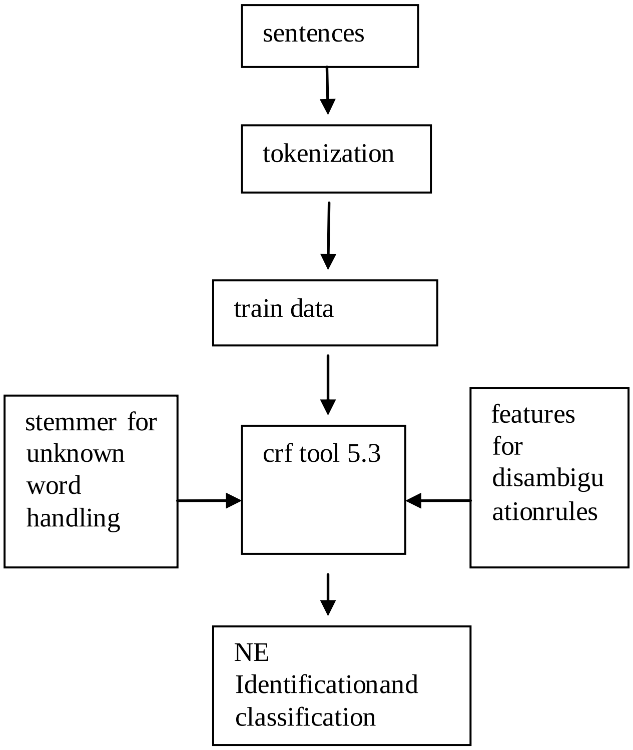


Fig 4.3.2 Functional Diagram for Named Entities Identification and Classification

The noun identification technique used is very useful for identification of 90-95% of nouns. This method identifies 90-92 % of proper nouns and resolves the ambiguity using good language independent rules.NE identification is the process of NER identification. It relies dictionary, suffix stemmer, features, pattern matching techniques, NE rules, disambiguation rules and context features. The process of NE identification reviews the input from the above process of Noun-identification in terms of nouns. NE identification, after reviewing the nouns, compares them with entries in NE Gazetteers containing a suitable application of NE features. Sometimes, this may result in ambiguity. These ambiguities are resolved in accordance with context features. Thus finally we get the desired NER output. Still, if any ambiguity is not resolved, then it is further supplied to a Machine Learning process for further resolution. The results of the Rule-based approach are used as training data for the machine learning approach. In the previous section ,we have used rule based approach to identify nouns and classify them. Here, we are going to use machine learning technique to identify and classify the named entities. The technique we are using are conditional random field algorithm. For the implementation of these Machines learning approaches large amount of annotated training data is required to train the NER system. In English, various gold standard annotated data is already openly available. Whereas in Indian languages, such kind of resources are not openly available. Machine learning techniques requires large amount of trained data which acquires large amount of memory. Due to this drawback, we are doing single token processing instead of multi token processing. Using machine learning techniques solely gives large amount of overhead. So, we are using both rule based approach and one of the machine learning technique i.e., condition Random Fields algorithm. In this phase, we are using morphological features for comparison. Morphological features are nothing but suffix lists, prefix lists, context features etc., If the word is found in standard lexicon, then a category will be assigned like noun (n), adverb (adv),adjective (adj), and verb (v). If a particular word is not found in the lexicon, check the word in NER gazetteers, namely, Person, Location, and Organization. Then the category will be assigned. If there no match with gazetteers, the word is checked with the use of pattern matching techniques for Person, Location, and Organization and the category will be assigned appropriately. If there is no match, then finally, using inflectional suffixes and feature suffixes (case markers), the category will be assigned. Every feature function in CRF has any real value on the basis of observation of the givenlanguage and these characteristic functions hold true for the whole model distribution too. First, the system requires some knowledge of the task of NE disambiguation. This representation is known as Language rules**.** Secondly, there is an NE disambiguation algorithm, which decides the best possible tag assignment for every word in a sentence according to the language model. The third component estimates the set of possible tags {T}, for every word in a sentence. This is referred to as possible classmodel.


sentences
tokenization
train data
stemmer for unknown word handling
crf tool 5.3
features for disambigu ationrules
NE Identificationand classification

This model consists of a list of lexical units associated with the list of possible tags. Here, it has been assumed that every word can be associated with any of the tags in the tag set i.e. a set of 4 tags in the tag set {T}). The input to the disambiguation algorithm takes the list of lexical units with the associated list of possible tags. The disambiguation module provides the output tag for each lexical unit using the encoded information from the language model .The last component, Unknown word handling, takes care of the words that are unknown to the system during training.We have used a freely available (open source) implementation of Conditional Random fields package.

5.EXPERIMENTAL RESULTS:

5.1 Input:

naku chaduvukovadam ante chaala istam . nenu chithoor lo nivasisthunnanu .nenutirupatilo s.v.university lo chaduvuthunnanu .nenu ma guidehumera khanammadam kinda projectchesthunnanu. ma kalasala udayam 8:30 kimodalavthundhi.

నాకుచదువుకోవడంఅంటేచాలాఇష్టం.నేనుచిత్తూరులోనివిసిస్తున్నాను.నేనుతిరుపతిలో,యస్.వి.యూనివర్సిటీలోచదువుతున్నాను. నేనుమాగైడ్హుమెరఖానమ్మేడమ్కిందప్రాజెక్టుచేస్తున్నాను. మాకళాశాలప్రొదున్న 8:30కిమొదలవుతుంది.

Step1: divide file into sentences

naku chaduvukovadam ante chaala istam.

నాకుచదువుకోవడంఅంటేచాలాఇష్టం.

nenu chithoor lo nivasisthunnanu.
నేనుచిత్తూరులోనివసిస్తున్నాను.

nenu tirupati lo s.v.university lo chaduvuthunnanu.
నేనుతిరుపతిలోయస్.వి.యూనివర్సిటీలోచదువుతున్నాను.

nenu ma guide humera khanam madam kinda project chesthunnanu.
నేనుమాగైడ్హుమెరఖానమ్మేడమ్కిందప్రాజెక్టుచేస్తున్నాను.

ma kalasala udayam 8:30 ki modalavthundhi.
మాకళాశాలఉదయం 8:30కిమొదలవుతుంది.

Step 2: Tokenization
naku | chaduvukovadam | ante | chaala | istam |
నాకు|చదువుకోవడం|అంటే|చాలా|ఇష్టం|

nenu | chithoor | lo | nivasisthunnanu |
నేను|చిత్తూరులో|నివసిస్తున్నాను|

nenu | tirupati | lo | s.v.university | lo | chaduvuthunnanu |
నేను|తిరుపతిలో|యస్.వి.యూనివర్సిటీలో|చదువుతున్నాను|

nenu | ma |guide | humera khanam | madam | kinda | project | chesthunnanu |
నేను|మా|గైడ్|హుమెరఖానమ్|మేడమ్|కింద|ప్రాజెక్టు|చేస్తున్నాను|

ma | kalasala | udayam | 8:30 | ki | modalavthundhi|
మా|కళాశాల|ఉదయం| 8:30కి|మొదలవుతుంది|

Step 3: If tokens directly match with dictionary, assignas noun
తిరుపతి (Tirupati) (noun)

హుమెరఖానమ్ (Humerakhanam) (noun)

Step 4: Otherwise, do stemming.
యస్వియూనివర్శిటిలో(s v university lo)

తిరుపతిలో(tirupati lo)

8:30కి (8:30 ki)

Step 5: If the noun match with NER list then assign itstag, otherwise use NER features and DisambiguationRules using rule-based approach.
తిరుపతి(tirupati) (Location name Vs Person name)

యస్వియూనివర్శిటి(svuniversity)(Organization name)

హుమెరఖానమ్(humera khanam) (Person name)

8:30(time)
Step 6: Still have ambiguity and unknown words, remove them using CRF in the stemmer
తిరుపతి (tirupati)(Location name)

5.2 Output:
తిరుపతి(tirupati) (Location name)

యస్వియూనివర్సిటీ (s v university)(Organization name)

హుమెరఖానమ్(humera khanam) (Person name)

8:30 (time)

5.3 Performance Measures

Our approach is evaluated with three major performance evaluations. The performance metrics used are Precision (P), Recall (R), and F-measure (F).The following measures are used for evaluation of our NER system.

Precision (P):

Precision is the part of the retrieved documents that are related to the user's information need and is the Ratio of Correct answers to the answers produced.

Precision (P) = correct answers / answers produced

Recall (R):

Recall is the fraction of the documents that are relevant to the query that is successfully retrieved and is the Ratio of Correct answers to the total correct possible answers.

Recall (R) = correct answers / total correct possible answers.

F-Measure (F):

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is

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

We have used suffix and context features and gazetteers for identification of proper nouns. These methods identify proper nouns correctly and resolve the ambiguity using good language independent rules or any statistical approaches give very good performance. The Individual approaches also do not give good results for Telugu. In future multi-handling can be done directly.

6. CONCLUSION AND FUTURE WORK

We have developed automatic NE identification and classification system using Rule based approach and CRF approach. We observed that Rule-based approach is language-dependent. Once rules are developed, large amounts of data can be processed for the identification of NEs. It is error-free, and there is no inconsistency in the data, since no manual training data is used. It is developed on purely language specific rules. This output is used as training data in machine learning approaches. Our Rule-based approaches are not fully automatic. It is a semi-automatic for handling unknown words and ambiguity words. Our approach is giving more than 92% accuracy on Telugu newspaper text. Our Rule-based system works only for individual tokens. Each token can be handled at a time in our process. Multi-token words cannot be handled directly. These words can be handled indirectly using context features. In future multi-token words list can handle these words directly.More accuracy needs more rules and more suffixes to the system. If we give more data and train them the we can improve the new models which thereby increase the accuracy of the system. Machine learning approach is fully automatic for identification and classification. Unknown words and ambiguity words are resolved in this approach. But it needs more training data for Machine learning. It gives better performance but cannot handle multi-token words directly. To do this some other features and techniques are needed. In future these features can handle special domain names, product names and other loan words.

In future work,using different techniques we can process more than one token or multi-tokens and increase the performance of the system.

REFERENCES

1. Alfonseca, Enrique and Manandhar, S. 2002. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In Proceedings of the International Conference on General Word net.
2. Andrew Borthwick. 1999. Maximum Entropy Approach to Named Entity Recognition Ph.D. thesis, New York University.
3. Bh.Krishna Murthy and J.P.L.Gywnn. 1985. A Grammar of Modern Telugu. Oxford University Press, Delhi.
4. Bikel, Daniel M., Schwartz, Richard L. and Weischedel, Ralph M., 1999. An Algorithm that Learns What's in a Name. Machine Learning, Volume (34), 211-231.
5. P Sindhusree and Dr.MHumeraKhanam ,Named Entity Recognizer for Telugu language using hybrid approach, International Journal on Recent and innovation trends in Computing and Communication. ISSN 2321:8169 Vol 4- Issue132-139
6. Brown, C.P., The Grammar of the Telugu Language. 1991, New Delhi: Laurier Books Ltd.
7. Collins, Michael and Y. Singer. 1999. Unsupervised models for Named Entity Classification , In proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.
8. Florian, R., luycheriah, A., Jing, H., Zhang, T. 2003. Named Entity Recognition through Classifier Combination. In Proceedings of the International Conference on Natural Language Learning (CoNLL-2003), 168-171, Edmonton,Canada.
9. Hideki Isozaki. 2001. Japanese named entity recognition based on a simple rule generator and decision tree learning. In proceedings of the Association for Computational Linguistics,