ISSN (Online): 2230-7893
www.IJCEM.org

Named Entity Recognition for Telugu Language

M. Humera Khanam[1], P. Udayasri[2]

[1,2]Dept. of Computer Science and Engineering,
SVU College of Engineering, Tirupati, Andhra Pradesh, India

Abstract

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify elements in text into predefined categories such as names of persons, locations, organizations, date, measures etc; NER has many applications in Natural Language Processing (NLP). Data classification, more accurate internet search engines, automatic indexing of documents, automatic question answering, cross language information access, and machine translation system etc are applications of NER in NLP.

This paper describes about the development of a two stage hybrid Named Entity Recognition system for Telugu language. We have used Maximum Entropy Model in this system. We have used variety of features and contextual information for predicting the various Named Entity (NE) classes. We have also tried to identify the nested named Entities by giving some linguistic rules.

Keywords : *Named Entity Recognition, Maximum entropy approach, NE features, Telugu*

1. Introduction

Natural Language processing refers to the use and ability of systems to process sentences in a natural language such as English, Telugu rather than in a specialized artificial computer language such as C, C++, java etc;[12]. The development of NLP applications is challenging because computers traditionally require humans to speak to them in computer programming language that is precise, unambiguous and highly structured commands. Human language, however, is not always precise. It is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

The ultimate goal of NLP is to do away with computer programming languages altogether. Instead of specialized languages such as Java or Ruby or C, there would only be "human".

Some of the tasks in Natural language Processing include Automatic summarization, Discourse analysis, Machine Translation, Morphological segmentation, Named Entity Recognition, Natural Language generation, Natural Language understanding, Parts of Speech tagging, Parsing, Question answering, Sentiment analysis, Speech recognition, Word sense disambiguation.

2. Approaches in Ner

Various approaches used for NER are Rule based approaches, machine learning techniques or statistical approaches. Both methods can be combined to yield best results.

2.1 Rule based approach

Rule based approach requires hand written rules which requires knowledge on specific language. In this approach rules are used to identify named entities. NER system uses gazetteer to classify words[4]. In this approach some language based rules and other heuristic are used to classify words. It needs rich and expressive rules and gives good results. It requires an advanced knowledge of grammar and other language related rules.

2.2 Machine learning based approach

Machine learning techniques uses large amount of annotated data to train the model. Several ML techniques include Hidden Markov models, Maximum entropy model, Conditional random fields and Support vector machines[10]. This approach explores the study and algorithms that can learn from and make predictions on data. This approach is used to build a model from example inputs in order to make predictions or decisions.

2.3 Hybrid approach

In Hybrid approach both rule based approach and machine learning approach is used to improve accuracy of a model. Some times more than one machine learning approaches are used in a model in order to improve accuracy. For example Hidden markov model and Maximum Entropy model can be used to design a model.

3. Design Challenges

NER is most challenging task in the field of NLP. Most of the text processing applications such as search systems, spelling checkers do not treat proper names correctly. This implies that names are difficult to identify and interpret in unstructured data. There are several

reasons for this difficulty. There are some rules like capitalization in English language to identify named entities. But in non-English languages like Indian languages, there exist no such rule.

In languages like Telugu, same name can be interpreted as person name and location name. We need to resolve this ambiguity. Most of the Indian languages do not have proper resources like corpus, dictionaries, gazetteer lists. Due to unavailability of resources, the task of identifying named entities in Telugu language is becoming challenging task. Detection of nested entities is also design challenge of NER. Most of the existing named entity recognition models focused only on single word entities. To identify nested entities we need to make some rules.

Ambiguity is major challenge. For example

- తిరుపతి can be identified as person name and location name.
  [tirupati] vs [tirupati]
- సత్యం can be part of person name and part of organization name.
  [satyam] vs [satyam]
- విశాఖపట్నం can be location name or organization name.
  [vishaakhapaTnam] vs [vishaakhapaTnam]

4. Proposed Method

NER for Telugu language is a two tier process. In first step, nouns are identified in given test file. In this step, Maximum Entropy approach is used for identifying nouns. In second step, Named Entities are identified. To find Named Entities several gazetteer lists and some rules are used. With this model the following named entities are identified.

- Person names
- Location names
- Organization names
- Date
- Time
- Week

4.1 Noun Identification

To identify noun phrases, one of the machine learning methods Maximum entropy model is used. Machine Learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt.

The principle of maximum entropy states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy. This model assigns outcome for each token in test data based on history and features. This model calculates probability values of a word belonging to each class. That is, given a sequence of words, the probability of each class is obtained for each word. To find the most probable tag corresponding to each word of a sequence, we can choose the tag having the highest class conditional probability value.

In this model, we use training corpus with only three tags, 'noun', 'verb', and for all words 'others' tag. We developed our own corpus. Data is collected from various resources like newspapers, blogs, Wikipedia.

Mathematical model for Maximum Entropy model: Let X be the set of conditions, and Y the set of possible outcomes.

$$P(Y \mid X)=1/Z(X)\exp([\Sigma\lambda_i f_i(X,Y)] \quad (1)$$

where λi –s are the parameters chosen to maximize the likelihood of the training data, and Z(x) is a normalization constant, which ensures that for every x the sum of probabilities of all possible outcomes is 1. The features used in the maximum entropy framework are binary. An example of a feature function is

f(X,Y) = { 1 if word(X)=kaaMgrees and type=org
0 otherwise

4.2 Named Entity identification

To identify Named Entities, Rule Based approach is used. A rule based systems needs more grammatical and linguistic analysis to make rules. We observed that Rule Based approaches may give good result with sufficient gazetteers lists, language dependent features and rules. These rules are language dependent. To identify Named entities like Date, Time we have used regular expressions.

Every language uses some specific patterns which may act as ending words in proper names and the list of this type of words is called as suffix list. For example

- వాడ(vaaDa), జిల్లా(jillaa), పట్టణం(paTTaNam), పల్లె(palle), పురం(puram) – suffix list for location names.
- నాయుడు(naayuDu), రెడ్డి(reDDi), శర్మ(sharma), రాజు(raaju), చౌదరి(choudari) – suffix list for person names
- యూనివర్సిటీ(yuunivarsiTee), సంస్థ(samstha), పార్టీ(paarTii) – suffix list for organizations name.

- To identify person names we can consider context features like శ్రీ(shree), మిస్టర్(misTar), మిస్(mis), డాక్టర్(DaakTar).

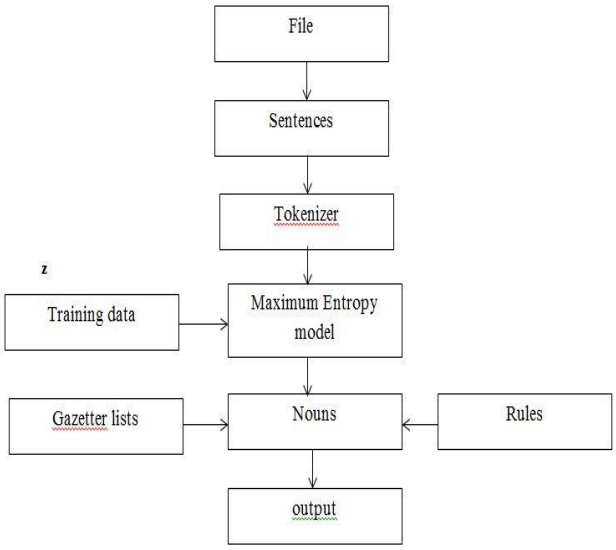
We need to make rules to identify Named Entities having ambiguity. For example

- If w_i is person name and w_{i+1} is preposition then assign w_i location tag else assign w_i person tag.
- If w_i is person name and w_{i+1} is a token from organization suffix list, then assign w_i organization tag else assign w_i person tag.

5. Algorithm

Step 1: Read test file and perform tokenization using delimiters "., ".
Step 2: Read training data that contains parts of speech (POS) tag for each word in training file.
Step 3: Test file and Training file are given as input to Maximum entropy model. Maximum Entropy model considers each POS tag in training file separate class.
Step 4: Find the probabilities of each class for every token in test file.
Step 5: Consider the class for each token having highest probability.
Step 6: Consider tokens with tags "noun" and "others", ignore rest of the tokens.
Step 7: Prepare gazetteer lists for each Named Entities separately and prepare gazetteer lists for suffix lists.
Step 8: Prepare rules to resolve ambiguity.
Step 9: Tokens with "noun" tag and "others" tag are compared with prepared gazetteer lists and rules.
Step 10: Assign Named Entity tags.

6. Block Diagram



7. Experimental Results

7.1 Sample Input test file:

నా పేరు ఉదయశ్రీ.ననేను 31-01-1993 వ తేదీన ఉదయం 10:00 AM కి జన్మించాను.ననేను యస్వియూనివర్సిటీ లో చదువుతున్నాను.ననేను ప్రొఫెసర్ హుమేరాఖానం గారి గైడెన్స్ లో ప్రొజెక్ట్ చేస్తున్నాను.ఉమ్మడి ఆంధ్రప్రదేశ్ రాజధాని హైదరాబాదు.హైదరాబాదు రంగారెడ్డి జిల్లా లో ఉంది.ఆంధ్రప్రదేశ్ ముఖ్యమంత్రి చంద్రబాబు నాయుడు గారు.ఆదివారం మా కళాశాల కు సెలవు.

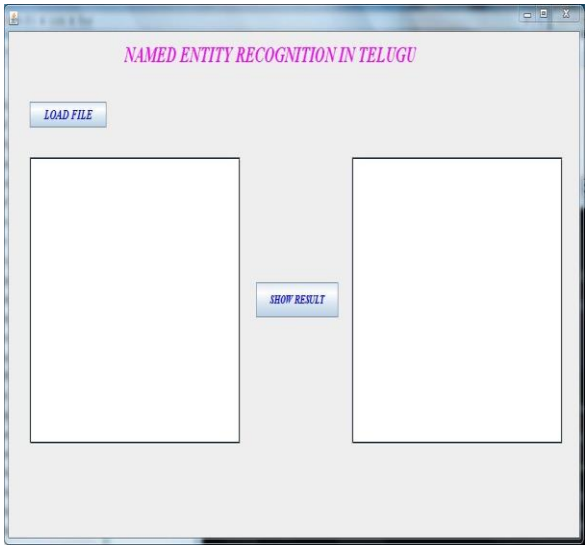
7.2 Transliterated form for input file:

naa pEru udayashree.nEnu 31-01-1993 va tEdeena udayam 10:00 AM ki janminchaanu.nEnu yasviyuunivarsiTy lO chaduvutunnaanu.nEnu profesar humeraakhaanam gaari gaiDens lO projekT chEstunnaanu.ummaDi aandhrapradEsh raajadhaani haidaraabaadu.haidaraabaadu rangaareDDi jillaa lO undi.aandhrapradEsh mukhyamantri chandrabaabu naayuDu gaaru.aadivaaram maa kaLaashaala ku selavu.

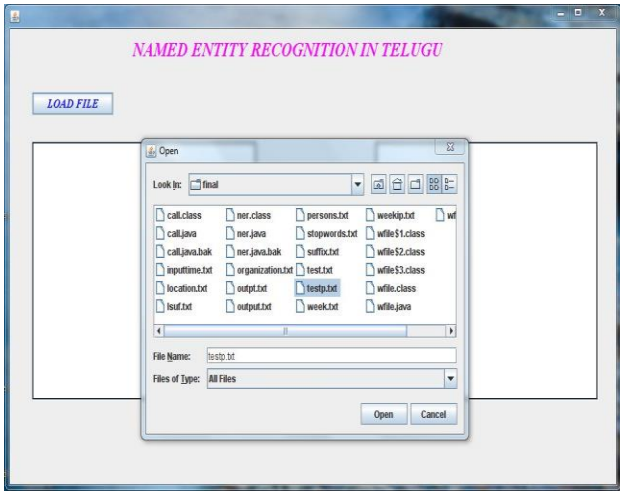
Output:

ఉదయశ్రీ(udayashree)	- person
31-01-1993	-date
10:00 AM	-Time
యస్వియూనివర్సిటీ(yasviyuunivarsiTy)	-organization
హుమేరాఖానం(humeraakhaanam)	-person
హైదరాబాదు(haidaraabaadu)	-location
రంగారెడ్డి(rangaareDDi)	-location
చంద్రబాబు నాయుడు(chandrabaabu naayuDu)	-person
ఆదివారం(aadivaaram)	-week

- GUI for NER in telugu



- When click on load file button



- Loading test file

- when click on show result button

Conclusion and Future Work

In this paper, we have presented NER system for Telugu. In this model we have used Hybrid approach i.e; combination of maximum entropy method and rule based approach. This model identifies nested entities also. We can try to identify more Named Entities. It works only for Telugu language.We can try to improve the model by making this language independent model.

References

[1] Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dantapat,Sudeshna Sarkar and Pabitra Mitra 2008 "A Hybrid Approach for Named Entity Recognition in Indian Languages" Proceedings of the IJNLP-08 workshop on Ner for South and South East Asian Languages.

[2] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Miguel Gómez-Berbís," Named Entity Recognition: Fallacies, Challenges and Opportunities".

[3] Ekbal A. and Bandyopadhyay S. 2007. Lexical Pattern Learning from Corpus Data for Named Entity Recognition.In Proceedings of International Conference on Natural Language Processing (ICON), 2007.

[4] Cucerzan Silviu and Yarowsky David. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC 1999, 90–99.

[5] Gu B. 2006. Recognizing Nested Named Entities in GENIA corpus. In Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 112-113.

[6] Kumar N. and Bhattacharyya Pushpak. 2006. Named Entity Recognition in Hindi using MEMM. In Technical Report, IIT Bombay, India.

[7] H. L Chieu., H Tou Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information", In Proceedings of the 6th Workshop on Very Large Corpora, 2002.

[8] A. McCallum, D. Freitag, F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation", In Proceedings. of the 17th International Conference on Machine Learning,2000, pp.591-598.

[9] GuoDong Zhou and Jian Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics, pages 473–480.

[10] A. Borthwick., "A Maximum Entropy Approach to Named Entity Recognition, Ph.D thesis, New yark University.

[11] Kumar, N. and Pushpak Bhattacharyya. 2006.Named Entity Recognition in Hindi using MEMM. Technical report, IIT Bombay, India.
[12] Srikanth, P, Murthy, Kavi Narayana,"Named Entity Recognition for Telugu",Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, 2008, IIIT, Hyderabad, India.
[13] B. Andrew, "A Maximum Entropy Approach to NER," Ph.D.dissertation, 1999.
[14] Michael Fleischman, "Automated sub categorization of named entities". Proc. Conference of the European Chapter of Association for Computational Linguistic, pp 25–30, 2001.
[15] Yungwei ding hsinhsi Chen and Shihchung TsaI, "Named entity extraction for information retrieval". Proc. of HLT-NAACL.
[16] G. Raju, B.Srinivasu, D. S. V. Raju, and K. Kumar, "Named Entity Recognition for Telugu using Maximum Entropy Model," Journal of Theoretical and Applied Information Technology, vol. 3, pp. 125–130, 2010.
[17] P.Sindhu sree, Dr. M. Humera Khanam, "Named Entity Recognizer for Telugu language using Hybrid approach", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 1,132-139

Authors

Dr. M. Humera Khanam had been awarded with PhD degree in Computer Science and Engineering from Sri Venkateswara University, Tirupati, A.P, India in the year 2015. She completed her M.Tech and B.Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University Anantapur, A.P,India in the years 2004 and 1998. She is now working as Associate Professor in the Department of Computer Science and Engineering, Sri Venkateswara University. She has 18 years of teaching experience. She is TPC member in IJNLP. She published more than 20 papers in reputed international journal with high impact factor. She attended 7 international conference and presented papers. She chaired as a resource person for AICTE sponsored seminars, workshops and conferences. Her areas of interests include Speech and Natural Language Processing, Machine Learning, Human Computer Interaction and Theory of Computation and Artificial Intelligence.

Ms. Udayasri received the B.Tech degree in 2014 from Jawaharlal Nehru Technological University, Anantapur, Andhra pradesh, India. Currently she is pursuing her M.Tech degree from Department of Computer Science and Engineering, Sri Venkateswara University, Tirupati, Andhrapradesh, India. Her current research interests include Machine Learning, NLP.