NAMED ENTITY RECOGNITION FOR TELUGU USING MAXIMUM ENTROPY MODEL

[1]G.V.S.RAJU,[2] B.SRINIVASU, [3]Dr.S.VISWANADHA RAJU,[4]K.S.M.V.KUMAR

[1] Professor, Department of Computer Science and Engineering, IIET, Siddipet, India -502277

[2]Asso.Prof.Department of Computer Science and Engineering, IIET, Siddipet, India -502277

[3]Professor,Department of Computer Science and Engineering, GRIET, Hyderabad, India -500072

[4]Asso.Prof.Department of Computer Science and Engineering, PPITES, Khammam, India -507305

E.Mails: letter2raju@gmail.com, srinivas_534@yahoo.com, viswanadharajugriet@gmail.com, ksmvkumar@yahoo.co.in

ABSTRACT

This paper describes a maximum entropy named entity(Identifying and classifying personal, location, organization or other names) recognition system for Telugu language. Named entity (N.E) recognition is a form of information extraction in which we seek to classify every word in a document as being a person-name, organization name, location name, date or none of the above. Named entity recognition is an important task for many NLP applications.

Keywords : *Named Entity Recognition, Maximum Entropy, Named Entity, NLP,Telugu*

1. INTRODUCTION

Named Entity Recognition(NER)(which might also be called as proper name classification and identification) is a computational linguistic task in which we seek to classify every word in a document as falling in to one of the four categories : person, location, organization ,and name-others(Date and time, etc..). In the taxonomy of the computational linguistic tasks, it falls under the domain of "information extraction". Information Extraction (IE) pulls facts and structured information from the content of large text collections.

There are many levels of sophistication which one can attempt in information extraction. The most ambiguous task currently being widely attempted is "scenario template" extraction. In this task, it is required to to retrieve a wide variety of information about certain type of event from a document.

The ability to determine the named entities in a text has been established as an important task for several natural language processing areas, including information retrieval, machine translation, information extraction and language understanding. NER for Indian Languages is a challenging task. There is not much work done in NER for Telugu in particular. Telugu, a language of the Dravidian family, is spoken mainly in southern part of India and ranks second among Indian languages in terms of number of speakers. Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. There are few languages in the world that match Telugu in this regard. the large number of morphological variants for a given root. High performance morphological analyzers have eluded researchers for a long time.[9]

2. NAMED ENTITY RECOGNITION

Named Entity Recognition, which is the subject of the paper, is much simpler than either of the tasks described above and it is a necessary precursor to them. Clearly, before we can determine the relationship between satyaM సత్యం (Satyam) and kooneeru haMpi కొనేరు హంపి (Koneru Humpy), we must first properly categorize them respectively as

an organization and a person. Similarly s`reeharikooTa శ్రీహరికోట(Sriharikota)must first be identified as a location before we can identify it as a satellite launch center.

Since the named entity task is very important sub-task of many applications like IE, MT, QA etc, a relatively high accuracy rate is expected of N.E systems. There are large numbers of ambiguous cases which make it difficult to attain human performance levels on the task. For instance:

When is the word ``tirupati" being used as the name of a person and when as the name of the city?

NER for Indian languages like Telugu is a very challenging task. Telugu like other Dravidian languages is a morphologically rich language. A Named Entity (NE) is a word or sequence of words that can be classified as a name of a person, organization, location,and not-name. In Information Extraction systems, accurate detection and classification of NEs is a very important task given that NEs can help us to extract knowledge from the texts; such as where the event happened, who were involved and when it happened. Named Entities are valuable in several NLP applications like

Machine Translation, Question and Answering systems, Automatic summarization and they are also very useful in the field of IR in building more accurate internet search engines.

3. APPLICATIONS OF NAMED ENTITY RECOGNITION

1. Indexing purposes in Information Retrieval General Document organization. A user can call up all the documents on company intranet which mention particular individual
    1. Before reading a article, user could see list of people, places and companies mentioned in the document.
    2. Automatic indexing of books. For many books, the majority of items which would go in the index would be named entities
    3. A named entity tagger can serve as a preprocessing step to simplify tasks such as machine translation.
    4. As mentioned earlier, an N.E tagger is an essential component of more complex IE tasks.
    5. Automatic summarization
    6. Question and Answering system.

4. BASIC PROBLEMS IN NAMED ENTITY RECOGNITION

Ambiguity in NE:

1. Tirupati(తిరుపతి (tirupati)) person name Vs place name
2. baMgraaru(gold) (బంగారు) person first name Vs common noun.
3. When is the word " prakaaSaM" ప్రకాశం being used as the name of a person and when as name of the city?
4. TATA (టాటా(TaaTaa)) person last name Vs organization.

Example TaTa Motors is a organization
Ratan TaTa is person name

Ambiguity with common words for example "raaju రాజు (king) and raaNi రాణి(queen)" can be a person name as well as a common word.

We have used part of news articles from iinaaDu (ఈనాడు) and vaarta(వార్త), Telugu wikipedia, Telugu local dailies for all our experiments. For all of our experiments we are using the roman form of these articles.

5. COMPLEXITY OF THE INDIAN LANGUAGES

NLP research around the world has taken giant leaps in the last decade with the advent of efficient machine learning algorithms and the creation of large annotated corpora for various languages. How ever NLP research in India have started with the development of rule based systems due to the lack of annotated corpora. Statistical NLP research in Indian languages can only be given a push by the creation of annotated corpus for Indian languages.

Among the modern Indian languages, Hindi has the simplest morphology and the complexity increases as we move from North to South India, with Dravidian languages having the most complex morphology.

6. RICHNESS IN MORPHOLOGY

Telugu, a language of the Dravidian family, is spoken mainly in southern part of India and ranks second among Indian languages in terms of number of speakers. Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and

statistical features [2], resulting in long and complex word forms. There are few languages in the world that match Telugu in this regard. Telugu is a free word order Language. The main reason for richness in morphology of Telugu (and other Dravidian languages) is, a significant part of grammar that is handled by syntax in English (and other similar languages) is handled with in morphology. Phrases including several words in English would be mapped on to a single word in Telugu. Thus "vacciveLLaaDu" వచ్చివెళ్ళాడు ((he) came and went),"veLLivastaaDaa" వెళ్ళివస్తాడు (will (he) went and come),"gelavaleeDanukonnaavaa" గెలవలేడనుకొన్నావా (do you think he will not win?), "raajamaMDrivaipu"రాజమండ్రివైపు(towards rajahm undary),"rajamaMDrinuMci" రాజమండ్రినుంచి(from rajahmundary) are all single words in Telugu, written and spoken as atomic units without spaces or pauses. Naturally we will see large number of types and the type token ratio is expected to be high. Telugu is primarily a suffixing Language an inflected word starts with a root and may have several suffixes added to the right. Suffixation is not a simple concatenation and morphology of the language is very complex.

NLP Applications such as IR and IE are often hard because of, among other things, the large number of morphological variants for a given root. High performance morphological analyzers have eluded researchers for a long time. Indian names are diverse, i.e there are lot of variations for a given named entity and also these entities are highly inflected such that none of the string matching algorithms work for getting the actual named entity. So until and unless we have good morphological analyzer and stemmer, it is not possible to identify the inflected entities automatically. For example ``bharatiiya janataa paarTii" భారతీయ జనతా పార్టీ(political party) is written as bi.je.pi (బి.జె.పి), baa.jaa.paa (బా.జా.పా), baajaapaa , etc. Different news articles have their own styles of writing. This makes the problem more difficult. Telugu is resource poor language. It lacks the proper amount of annotated data, name dictionaries, good morphological analyzers, parsers etc. Web resources for the name lists are found in English, but Telugu place names available on Telugu Wikipedia. So there is need for transliteration. Capitalization feature concept doesn't work for Indian languages, but it works well for English.

7. CORPUS

We have used part of news articles from iinaaDu and vaarta(ఈనాడు, వార్త)Telugu Wikipedia, Telugu local dailies for all our experiments. For all of our experiments we are using the roman form of these articles.

8. PERFORMANCE EVALUATION

Evaluation Metric mathematically defines how to measure the system's performance against human-annotated, gold standard.

Here for every experiment before checking the performance of the system, a human tagged test data is prepared to evaluate the system.
The system's performance is measured in terms of precision (P), recall (R) and f-measure (F).

P = {Correct-answers} /{answers-produced}

R = Correct-answers} {/Total-possible-correct-answers}

$F= (\beta^2 + 1)P R /(\beta^2 R + P)$

The typical value of β is 1.

9. A MAXIMUM ENTROPY APPROACH

The maximum entropy framework estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. Such constraints are derived from training data, expressing some relationship between features and outcome. The probability distribution that satisfies the above property is the one with the highest entropy. It is unique, agrees with the maximum-likelihood distribution, and has the exponential form [2]

$$\mathbf{p(o|h)} = \frac{1}{(Z(h))} \prod_{j=1}^{k} \alpha_j^{fj}(h,0)$$

where o refers to the outcojme, h the history (or context),and Z(h) is a normalization function. The features used in the maximum entropy framework are binary. An example of a feature function is

f(h,o) = { 1 if word(h)=kaaMgrees and type=org
0 otherwise

The parameters αj are estimated by a procedure called Generalized Iterative Scaling (GIS) (Darroch and Ratcliff, 1972). This is an iterative procedure that improves the estimation of the parameters at each iteration.

The maximum entropy classifier is used to classify each word as one of the following: the beginning of a NE (B tag), a word inside a NE (C tag), the last word of a NE (L tag), or the unique word in a NE (U tag).

During testing, it is possible that the classifier produces a sequence of inadmissible classes (e.g., PER-B followed by LOC-L). To eliminate such sequences, we define a transition probability between word classes P (ci |cj)to be equal to 1 if the sequence is admissible, and 0 otherwise. The probability of the classes c1 , . . . , cn assigned to the words in a sentence s in a document D is defined as follows.

$$P(c1,..., cn \mid s, D) = \prod_{i=1} P(ci|s, D) \times P(ci|ci-1)$$

where, P(ci | s, D) is determined by the maximum entropy classifier. The Beam search algorithm is then used to select the sequence of word classes with the highest probability.

The features are binary valued functions which associate a NE tag with various elements of the context.

10. FEATURE REPRESENTATION

LISTS DERIVED FROM TRAINING DATA

The training data is first preprocessed to compile a number of lists that are used by both ME1 and ME2. These lists are derived automatically from the training data.

Context word feature: Preceding and following words of a particular word.

Word suffix /prefix: Various word suffixes/prefixes can be used as the features in two different ways. The first one is to use a fixed length word suffix/prefix of the current and/or the surrounding word(s) as the features. If the length of the corresponding word is less than or equal to n-1 then

Frequent Word List (FWL) This list consists of words that occur in more than 5 different documents.

Useful Unigrams (UNI) For each name class, words that precede the name class are ranked using correlation metric (Chieu and Ng, 2002a), and the top 20 are compiled into a list.

Useful Bigrams (UBI) This list consists of bigrams of words that precede a name class. Examples are “graamaM loo”, “amalapuraM nuMci”, etc. The list is compiled by taking bigrams with higher probability to appear before a name class than the unigram itself. A list is collected for each name class. A possible explanation is that in writing,people tend to explain with bigrams such as “graamaM loo” before mentioning the name itself.

Useful Word Suffixes (SUF) For each word in a name class, final suffixes with high correlation metric score are collected. where suffixes such as “Reddy” and “naayuDu” often appear.

Useful Name Class Suffixes (NCS) A suffix list is compiled for each name class. These lists capture tokens that frequently terminate a particular name class. For example, the org class often terminates with tokens such as paarTii and saMstha(పార్టి, సంస్థ) , etc.

11. Gazetteer lists for Telugu: Various gazetteer lists have been developed either manually or semi automatically from the Telugu news corpus and Wikipedia. These lists have been used as the binary valued features of the ME model. Any particular list does not include the ambiguous entries, i.e., those that can appear in more than one gazetteer list. If the current token is in a particular list, then the corresponding feature is set to 1 for the current and/or the surrounding word(s); otherwise, it is set to 0.Following is the set of gazetteers along with the number of entries:

Organization clue word (e.g., saMstha, parTii(సంస్ల,పార్టీ),limited etc):54, Person prefixes (e.g., reDDy,murthi,naayuDu(రెడ్డి, మూర్తి,శర్మ,నాయుడు),etc.): 100, Middle names: 700, Surnames: 3270, Common location (e.g., jilla,graamaM, road(జిల్లా,గ్రామం, రోడ్డు etc.): 230, Designation words (e.g., adyaksuDu,maMtri(అధ్యక్షుడు, మంత్రి), ect.): 642, First names: 4,890, Location names: 19000, Organization names: 500, Month name (English and Telugu calendars): 24, Weekdays (English and Telugu calendars): 14, Measurement clue words: 52 entries.

12. EXPERIMENTS RESULT

The Telugu training and test data are part of the eenadu,vaartha news papers and , Telugu Wikipedia Corpus, Results in Table 1 are obtained by applying ME1, without the help of name lists, on the Telugu 2 languages.

The best results for Telugu are obtained using ME2, which made use of name lists and organization list compiled from the Telugu Wikipedia and the list provided with the training set .

Table 1. Number of Entities in Test Data Sets

Corpus	NPER	NLOC	NORG	NOTH
EE-1	321	177	235	6,411
EE-2	325	144	187	5221

Table 2. Number of Entities in Training Data Sets

Corpus	NPER	NLOC	NORG	NOTH
TR-1	804	505	235	16,411
TR-2	1325	744	287	25221
TR-3	2500	1400	679	50,525

Table 3: Performance of maximum entropy based named entity system
corpus taken from on line news papers

	Precision(P)	Recall (R)	F β =1
NPER	79.45%	65.95%	72.07%
NORG	80.86%	48.67%	60.76%
NLOC	71.08%	65.90%	68.40%
NOTH	72.20%	32.90%	45.28

NPER : Persion name ,,NORG : Oranization Name,
NLOC: Location Name , NOTH :Other Names

13. CONCLUSION :

In this paper, we have presented a NER system in Telugu langaguge with the maximum entropy(ME) frame work.this system makes use of the different contextual informationof the words along with different orthographic word level features. all features are language specific features that only apply Telugu language only.

REFERENCES

[1]. Daniel M. Bikel, R. Schwartz, Ralph M. Weischedel, "An Algorithm that Learns What's in Name", Machine Learning (Special Issue on NLP), 1999, pp. 1-20.

[2]. 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.

[3]. 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.

[4]. Andrew Borthwick, J. Sterling, E. Agichtein, R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7", MUC-7, 1998, Fairfax, Virginia.

[5]. J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", In Proc. of the 18th ICML, 2001, pp.282-289.

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

[7]. Xavier Carreras, Lluis Marquez, and Lluis Padro. 2002. Named Entity Extraction using AdaBoost. In Proceedings of the Sixth Conference on Natural Language Learning, pages 167–170.

[8]. Mohammad Hasanuzzaman, Asif Ekbal and S. Bandyopadhyay, "Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi ",International Journal of Recent Trends in Engineering, Vol. 1,No.1.

[10]. 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.

[11]. Saha, Sujan Kumar and Sarkar, Sudeshna and Mitra, Pabitra,"Feature selection techniques for maximum entropy based biomedical named entity recognition",journal of Biomedical Informatics, volume-42, number-5, 2009, 905-911.

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

[13]. Zhang l Le, "Maximum Entropy Modeling Toolkit for Python and C++ . http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

[14]. Tsuruoka, Yoshimasa; Tsujii, J. 2003. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In Proc. Conference of Association for Computational Linguistics, Natural Language Processing in Biomedicine.

[15]. Tjong Kim Sang, Erik. F. 2002. Introduction to the CoNLL-2002 Shared Task: Language Independent Named Entity Recognition. In Proc. Conference on Natural Language Learning.

[16].Tjong Kim Sang, Erik. F.; De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proc. Conference on Natural Language Learning.

[17]. Munro, Robert, Daren Ler, and Jon Patrick. 2003. Meta-learning Orthographic and Contextual Models for Language Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 , pages 192–195. Association for Computational Linguistics

[18]. Ekbal, Asif, R Haque, and S. Bandyopadhyay. 2008. Named Entity Recognition in Bengali: A Conditional Random Field Approach. In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP08), pages 589–594.

[19]. Ekbal, A. and S. Bandyopadhyay. 2007b. Pattern Based Bootstrapping Method for Named Entity Recognition. In Proceedings of the 6th International Conference on Advances in Pattern Recognition (ICAPR), pages 349–355. World Scientific.

[20]. Ekbal, Asif and S. Bandyopadhyay. 2008a. Bengali Named Entity Recognition using Support Vector Machine. In Proceedings of NERSSEAL, IJCNLP-08 , pages 51–58.

[21]. Ekbal, A. and S. Bandyopadhyay. 2008b. A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal 42(2):173–182.

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

[23]. Malouf, R. 2002. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In Proceedings of Sixth Conference on Natural Language Learning, pages49–55.