Rule based Methodology for Recognition of Kannada Named Entities

Bhuvaneshwari C Melinamath
Department of Computer and Information Science
University of Hyderabad,Hyderabad,India

Abstract- Named Entity Recognition (NER) is an important task in Natural Language Processing (NLP) applications like Information Extraction, Question Answering etc. In this paper we have proposed a rule based methodology to recognize Kannada named entities like person name, location name, organization name, number, measurement, time. We have manually developed suffix, prefix list and proper noun dictionary of 5000 words. Capitalization feature in English is useful for identifying named entities. But in Kannada capitalization feature is does not exists thus the NER task in Kannada becomes complex than in English. In Kannada proper nouns are indistinguishable from forms that are common nouns and adjectives. This ambiguity makes Kannada NER a challenging. The results of recognition are encouraging and the methodology has the precision around 86\%. Famous Kannada news paper Prajavani corpus is used to carry out experiments. The tool is language independent as there is no hard coding can be used for other Dravidian language group.

Keywords: Named Entity Recognition (NER), Natural Language Processing (NLP), Message Understanding (MU) , Information Extraction (IR).

I. INTRODUCTION

Named Entity Recognition (NER) is a task of finding and classifying proper name, location name and organization name, etc. in a text. Named Entity Recognition (NER) is an important task in many Natural Language Processing (NLP) applications like machine translation, question-answering systems and indexing for information or information Retrieval, data classification and automatic summarization etc. The first and most important subtask of Information Extraction (IE) is NER. Because of the importance of Information extraction (IE), DARPA initiated a number Massage Understanding Conferences (MUC) in the mid nineties like (MUC) Conference. According to the specification defined by MUC, the NER tasks generally work on six types of named entities like person name, location name, organization name, time, measurement and number. For example consider a sentence: Mr. Rahul joined as manager in Municipal Corporation in Bangalore on Monday 14 September 2004. The various named entities in this sentence are "Mr. Rahul is person entity", "Municipal corporation is organization entity", "Bangalore is location entity", "Monday September 2004 is time entity".

The approaches of named entity recognition are namely rule based and Machine Learning. Machine learning includes conditional random field, support vector machine (SVM), Maximum Entropy Model (MaxEnt), Decision Tree, Support Vector Machines HMM etc. All the approaches make use of gazetteer information to build system, because it improves the accuracy. Modern systems most often use machine learning techniques since rule based approaches need months of development by experienced linguists whereas machine learning techniques uses collection of annotated documents to train classifier, however handcrafted rule-based systems usually give good results. But the main disadvantages of these rule- based techniques are that these require huge experience and grammatical knowledge of particular languages.

We have developed a rule based approach for named entity recognition, we have manually prepared proper noun dictionary of 5000 words, suffix, prefix gazetteer information and set of handcrafted rules derived from knowledge. Evaluation of our approach shows that it performs better and hence is easier to deploy for practical use.

The remaining part of the paper comprises of six sections. Section 2 gives description of Kannada language, section 3 describes the works in this area, section 4 discusses challenges in Kannada language, section 5 deals with our proposed method, section 6 gives results and discussions and section 7 deals with conclusion.

II. DESCRIPTION OF KANNADA LANGUAGE

Dravidian languages have a history of more than 2,000 years. Kannada is a Dravidian language spoken mainly in southern part of India and ranks third among Indian languages in terms of number of speakers as notified in census information. Kannada 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 Kannada in this regard. Kannada has productive system of derivation, saMdhi and compounding.

A single root can lead to the formation of a very large number of surface word forms. Words in Dravidian languages in general and (Kannada in particular) are an order of magnitude more complex than those in Indo-Aryan languages. The main reason for richness in morphology of Kannada and other Dravidian languages is, a significant part of grammar that is handled by syntax in English and other similar languages, is handled within word morphology in Kannada, for example the word 'baravudillavenu' in Kannada is eqivalentto several words (that is, tokens) in English like (do you think he/she/they/it will not come?). Kannada is a language of Dravidian family. Kannada language uses 49 phonemic letters, divided into 3 groups: swaragaLu "vowels" 13 in number, vyaNjangaLu " consonants" 34 in number and yogavaahakagaLu (neither consonant nor vowel two in number: anusvara “aM ", visarga “ah ").

Table -1 Consonants In Kannada

ಕ	ಖ	ಗ	ಘ	ಙ
ka	kha	ga	gha	n`a
ಚ	ಛ	ಜ	ಝ	ಞ
ca	cha	ja	jha	N`a
ಟ	ಠ	ಡ	ಢ	ಣ
Ta	Tha	Da	Dha	Na
ತ	ಥ	ದ	ಧ	ನ
ta	tha	da	dha	na

ಪ	ಫ	ಬ	ಭ	ಮ	ಯ	ರ	ಲ	ಳ	ವ	ಶ	ಷ	ಸ	ಹ
pa	pha	ba	bha	ma	ya	ra	la	La	va	s`a	sha	sa	ha

.

Table -2 Vowels In Kannada

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ
a	aa	i	ii	u	uu	R	R	e	ee	ai	o	oo	au

Anusvaara (Bindu or Sonne) ಂ aM, Visarga ಃ ah.

III. LITERATURE SURVEY

NER has drawn more attention from NLP researchers since the last decade. Lots of work is done on NER for English employing the machine learning techniques, using both supervised learning and unsupervised learning. Ralph Grishman in 1995 developed a rule based NER systems which uses some specialized name dictionaries including names of all countries, names of major cities, names of companies, common first names etc. and reports the recall, precision and f-measure as 86%, 90% and 88.19 % respectively (Grish, 1995). Borthwick in 1999 developed a ML based system i.e. MaxEnt based system, the system used 8 dictionaries (Borth, 1999). Fleischman proposed a method for categorization of location names using Bayesian and decision tree and the accuracy is 80% (Fleischman, 2001). Hsin Chen et.al (Hsin, 2003). proposed an algorithm for Named entity for Information retrieval, different types of information from different levels of text are employed, including character

conditions, statistic information, the recall rates and the precision rates for the extraction of person names, organization names, and location names are (87.33%, 82.33%), (76.67%, 79.33%)and (77.00%, 82.00%), respectively. M collins (Collin, 2002) proposed ranking algorithms for Named Entity Extraction, using maximum entropy and reports precision recall and F-measure as 84%,86%,85% respectively.

Shilpi Srivastava et. al proposed a hybrid approach, a combination of rule based CRF and maxEnt for Named entity recognition system for Hindi Language reports precision 96%, recall 86.96% and f-measure is 91% (shilpi, 2011). Asif Exbal proposed CRF based approach for Bengali and reports F-score of 89.3 % (Ekbal4, 2008). Riaz proposed a rule based approach for Urdu using small scale gazetteers and reports recall 90%, F-measure 93.14% and precision 96.4% (Riaz ,2010).

Not much work has been done on NER for Indian languages like Kannada. Even though Kannada is the most spoken language it lacks behind in terms computational technology and still there is no accurate NER system exists for Kannada. As some features like lack capitalization, lack of a large labeled dataset and lack of standardization and spelling variations makes Kannada NER a difficult task. Moreover English NER system cannot be used directly for Kannada. Hence there is a need to develop an accurate Kannada NER system for better presence of Kannada in the field of NLP on the world wide web. It is necessary to understand Kannada language structure and learn new features for building better Kannada NER systems. The importance of Ner system in NLP application is of higher significance and it is necessary to have such tool for Kannada.

From the literature survey it is clear that statistical methods require large annotated corpora, as annotated corpora is not available for Kannada, rule based approaches are the best alternative, so we are proposing the rule based methodology for Kannada NER here. As there is not much work reported for Kannada towards named entity recognition using rule based approach, that is also another motivation to move in this direction.

IV. DESIGNING NER IS COMPLEX FOR KANNADA

The certain features in Kannada like lack of capitalization Kannada, non-availability of large gazetteer, lack of standardization and spelling, scarcity of resources and tools, free word order of language, agglutinative property, and lack of web sources for name lists makes the named entity recognition a difficult task.

Owing to the above mentioned complexities inherent to Kannada makes designing NER a challenging task. We have proposed a methodology for NER taking in to consideration varieties of usages of location, organization, time and measurement information.

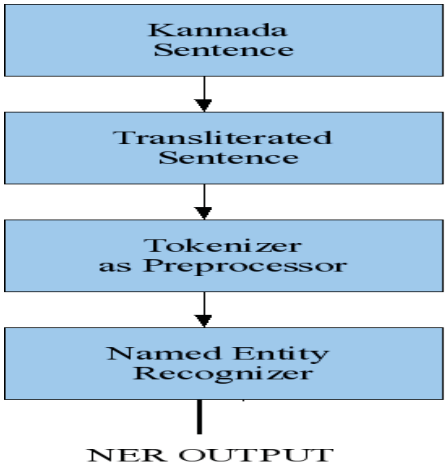


Figure 1. Named Entity Recognition System

V. PROPOSED METHOD

The architecture of proposed automatic NER system is as shown below. The main aim here is to design a system which takes a Kannada sentence as input and identifies and categorize named entities in the input. The design process for the system involves the following tasks: 1. Read transliterated file2. Tokenization module. 3. Dictionary lookup. 4. NER Module.

B. Transliteration–

First the raw corpus is converted in to transliterated corpus by using converter program and is given as input. The Unicode text file in Kannada font is converted to romanized or transliterated file intermediate map file is used for conversion between and English text, we have both Iscii to romanized conversion file vice-versa.

The operation of channel separation is applied on the watermarked color image to generate its sub images, and then 2-level discrete wavelet transform is applied on the sub images to generate the approximate coefficients and detail coefficients.

ಇವತ್ತು ನವರಾತ್ರಿಯ ಮೂರನೆಯ ದಿನ.
ಮೈಸೂರು ಮತ್ತು ತಂಜಾವೂರು ಎರಡೂ ೧೮-೧೯ ನೆ ಶತಮಾನಗಳಲ್ಲಿ ದಕ್ಷಿಣ ಭಾರತದ
ಪ್ರಮುಖ ಸಾಂಸ್ಕೃತಿಕ ನೆಲೆಗಳಾಗಿ ರೂಪುಗೊಂಡವು. ಹಾಗಾಗಿಯೇ ಇಂದಿಗೂ ನಾವು ವೀಣೆ,
ಚಿತ್ರಕಲೆ ಮತ್ತು ಭರತನಾಟ್ಯ ಇವೆರಡರಲ್ಲೂ, ಮೈಸೂರು ಶೈಲಿ ಮತ್ತು ತಂಜಾವೂರು ಶೈಲಿ
ಎಂದು ಎರಡು ಪ್ರಮುಖ ಶೈಲಿಗಳನ್ನು ನೋಡಬಹುದು. ಕರ್ನಾಟಕ ಸಂಗೀತ ತ್ರಿಮೂರ್ತಿಗಳೆಂದೇ
ಹೆಸರಾದ ತ್ಯಾಗರಾಜ, ಮುತ್ತುಸ್ವಾಮಿ ದೀಕ್ಷಿತ ಮತ್ತು ಶಾಮಾಶಾಸ್ತ್ರಿ ಇವರು ಬಾಳಿದ್ದು
ತಂಜಾವೂರು ರಾಜ್ಯದಲ್ಲೇ. ಆದರಲ್ಲಿಯೂ ಶಾಮಾಶಾಸ್ತ್ರಿ ಇವರು ಇದ್ದದ್ದಂತೂ ತಂಜಾವೂರು ನಗರದಲ್ಲೇ.

Figure 2 Sample Kannada Input File

English File of above Kannada Input file:
Today navaratri's third day. Mysore and TaNjavora two both 18-19 centuries south India's famous cultural places transformed. Hence today also we viiNe, painting art and bharatanatya in both of two, Mysore style and TaNjavora style two styles can be seen. Karnataka music three people famous names Tyagaraj, Muttuswami diikshita and Syamasyastri lived in TaNjavora state only. In that also Syamasyastri stayed in TaNjavora city only

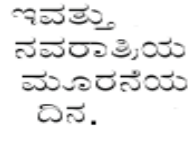
Figure 3 English Version of Kannada Input File in figure 2.

Transliterated File of above Kannada Input file:
ivattu navaraatriya muuraneya dina. maisuuru mattu taMjaavuuru eraDuu 18-19 ne s`atamaanagaLalli dakshiNa bhaaratada pramukha saaMskRtika nelegaLaagi ruupugoMDavu. Haagaagiyee iMdiguu naavu viiNe, citrakale mattu bharatanaaTya iveraDaralluu, maisuuru s`aili mattu taMjaavuuru s`aili eMdu eraDu pramukha s`ailegaLannu nooDabahudu. karnaaTaka saMgiita trimuurtigaLeMdee hesaraada tyaagaraaja, muttusvaami diikshita mattu s`aamaas`aastri avaru baaLiddu taMjaavuuru raajyadallee. adaralliyuu s`aamaas`aastri avaru iddaddaMtuu taMjaavuuru nagaradallee

Figure 4. Transliterated File of above Kannada Input File

C. Tokenization –

Tokenization process divides the text file in to sequence of tokens. And removes the delimiters like [.! ?:;] which are not part of words. We have also formulated some rules to identify abbreviations in the text. We listed some abbreviations and if the sentence boundary delimiter comes with this abbreviation; it may or may not be the exact sentence boundary.



etc. are tokens

D.NER Recognizer –

The named entity recognizer takes an input like "Ram" and output Ram as person name. The system uses set of pre -defined rules, suffix and prefix list, gazetteer list and dictionary.

E.Dictionary –

Each token is searched in the dictionary, if is found in the dictionary then tag corresponding to that is assigned to that token. If word is not found then it is passed to the NER module for further processing. We have developed a proper noun electronic dictionary of 5000 plus words manually for this work.

```
naDu||N-COM-UNC-N.SL-NOM::real -
u||ADJ-ABS
soodarasose||N-COM-COU-F.SL-
NOM::TYPE-kinship
ramaa||N-PRP-PERSN
siMdagi||N-PRP-LOC
```

Figure 5. Sample Dictionary

F. Features for NER –

Feature selection plays an important role in named entity recognition, different trial combinations are experimented to pick up suitable and all possible possibilities. The suffix and prefix information works well for highly inflected language like Kannada. Sliding window size feature, F= {wi-1, wi, wi +1, wi+ 2 etc.} We have manually prepared various gazetteer lists for use in NER like 11 location suffixes, 49 designation prefixes, 74 organization prefixes, 50 person prefixes, 32 measurement prefixes, 32 next word clues, and 5000 words proper-noun dictionary.

Following features are most often used for the recognition and classification of named entities.

I. Context word feature: Previous and next words of a particular word have been used as a feature.
II. Dictionaries: Dictionaries are used, Prefix and suffix lists were also important.
III. Named Entity Information: It is the features in which NE tag, tag of the previous word is considered. It is the dynamic feature.
IV. Gazetteer Lists: Due to the scarcity of resources in electronic format for Kannada language, so the gazetteer lists are prepared manually. Seven different lists are prepared.
V. Word suffix and prefix: Word suffix information is useful to identify NE's. Variable length suffixes of a word can be matched with predefined lists of useful suffixes for different classes of NEs. The different suffixes for person names are *gauDa, appa, swami,* etc., and for location names *baad, pura, haLLi* etc. And for long matching nested NEs *like akkamaadeevi raste*, is tagged as location NE instead as proper-noun

followed common noun. In this feature, a length of 1 to 4 words of the current and/or the surrounding words is taken.

VI. **Contains digit**: This features is useful in date kind of expressions, numbers, time expressions all formats of time like 8.00 AM/am, named entity measurement like 10 kg. date expressions like 12/09/2009,12-09-12,13-04-2012, 14 June 2009, Monday 14 January representing floating point values like 12,345. The features like digits and percentage, digits and hyphen, digits and period, digits and slash are handled.

VII. **Word clue:** This feature helps in identifying next or previous token as nes. For example raamaneMba huDuga 'boy called raama', words derived as past relative by adding 'aad' to all the animate words representing human beings like avara magaLaada meenaka 'their daughter named meenaka', the word following participle so formed is surely be a person representing named entity.

VIII. **Another cue is** , if token pattern is of the form Delhli.9 then the token preceding 9 is location named entity.

Table 3: Tag Sets Used for NER

TAG	DESCRIPTION	EXAMPLE
N-PRP-PRSN	Person Name	**Ram**
N-PRP-ORG	Organization Name	**Basaveshwar sugar ltd.**
N-PRP-LOC	Place Name	**Bombay**
N-PRP-NUM	Numeric Value	**12,345**
N-PRP-TIM	Time	**14 june Monday**
N-PRP-MEA	Measurement	**10 kg.**

VI. PROPOSED ALGORITHM

Algorithm for Named Entity Recognition and Classification

```
Input:  File containing transliterated Kannada sentences.
Output: Named Entities With Their Classes(Person,Organization,
        Location, Time, and Measurement);

Preprocess of text by splitting on space, removing unwanted charters like? -
Etc. and store in input file, if blank line do not process it.

Description: Wi-Current word, Wi+1 -next word, Wi+2-next (next (word), Wi-1-
previous word and soon.

Start
Read input File
  While (not end- of file) do
  R1: if(Wi is number)AND (Wi+1 is a Unit of measure)
      THEN (Wi, Wi+1) bigram is NE-Measurement

  R2: if (Wi is number) AND (Wi+1 is an am|pm Unit)
       THEN (Wi, Wi+1 bigram) is a NE-time

  R3: if (Wi is number) AND (Wi+1 is month name) AND (Wi+2 is
      digit no.)
      THEN (Wi, Wi+1, Wi+2) trigram is NE-Time

  R4: if (Wi is number) AND (Wi+1 is a month name) AND Wi+2 is
      not four digit no)
      THEN (Wi, Wi+1) bigram is a Time NE
  R5: if (Wi denotes day) AND (Wi+1 is a number) AND (Wi+2 is
      month name)
      THEN (Wi, Wi+1, Wi+2) trigram is NE-Time

  R6: if( Wi is day) AND (Wi+1 is a number) AND (Wi+2 is not
      month name )
```

```
      THEN (Wi,Wi+1,) bigram is a NE-time

  R7: if( Wi is day) AND (Wi+1 is not number) AND (Wi+2 is not
      month name)
      THEN (Wi unigram is a  NE-time )

  R8: if(Wi  matches with   month list) THEN Wi is  NE-Time

  R9: if(Wi matches date patterns) THEN(Wi)unigram is a time NE

  R10: if (Wi denotes suffix in the Designation list)
       THEN (Wi, Wi+1) bigram is NE-person

  R11: if(Wi ends with suffix in the person clue list) THEN Wi is
       NE-person
  R12: if (Wi ends with suffix in the location list) THEN Wi NE-
       location
  R13: if(Wi denotes suffix organization list)
        THEN (Wi, Wi-1, Wi-2) trigram NE-organization

  R14: if (Wi  matches  with suffix location street endings list)
       THEN (Wi,Wi-1) bigram is a NE-location

  R15: if (Wi mathces pattern character followed by dot followed
       by number)
       THEN Wi is NE-location or NE-Time based on pattern

#Search in Proper noun Dictionary
#Divide dictionary fields as Word and tag, store in hash;

R16: if (Search Wi the dictionary)
     THEN (Wi, Wi+1) bigram as NE-Location or NE-Organization.
     Otherwise Wi is NE-tag in dictionary

Read the input file
Stop
```

V II. RESULTS

To evaluate the NER system we randomly downloaded corpus from famous Kannada News paper PrajavaaNi website around 20 files of varying length and ran our NER through it, we observed that around 79.52% NEs are identified correctly, 12.34% are identified wrong and around 8.56% are left unrecognized. Precision is the percentage of correct positive predictions returned by the system. It is computed as the ratio between the number of NEs correctly identified by the system True Positives (TP) and the total number of NEs returned by the system. The precision is calculated by dividing TP by the sum of TP and false positives (FP).

$$precision = \frac{Tp}{Tp+Fp}$$

TP= 567, TP+FP=567+88. So Precision is 86% . Recall indicates the percentage of positive cases recognized by the system. It is computed as the ratio between the number of NEs correctly identified by the system (TP) and the number of NEs that the system was expected to recognize. Thus, Recall is the number of (TP) divided by the sum of (TP) and false negatives (FN).

$$\mathrm{Re}call = \frac{TP}{TP+FN} \qquad \text{and} \qquad F = (\beta+1) * \mathrm{Pr}ecision * \mathrm{Re}call \Big/ \beta(\mathrm{Pr}ecision + \mathrm{Re}call)$$

TP=567,TP+FN=567+61. Recall is 90%. F-measure is the common weighted harmonic mean between Precision and Recall defined as where beta is the weighting factor. When the Precision and Recall have the same values.

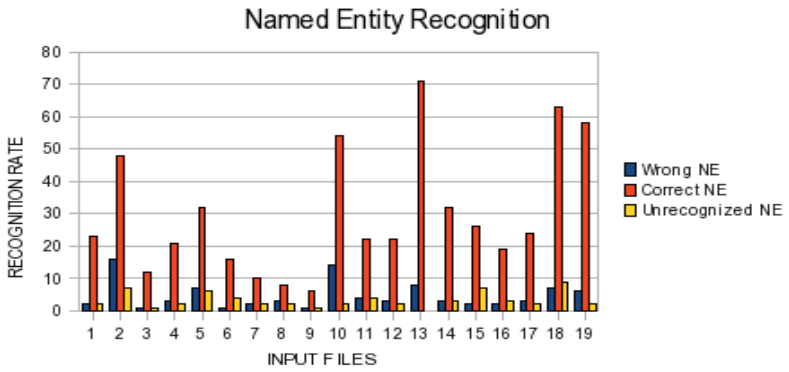
Table 4: Evaluation of NER Output

Number of Total NEs in Test Data	713
Number of NEs identified	655
Number of Correct identified NEs	567
Precision	86%
Recall	90%
F-measure	87.95%

Table 5: Sample Output of File in Figure 2.

S.No	English Word	Trasliterated word	Tag
1	Mysore	maisuuru	N-PRP-LOC
2	Tanjavore	taMjaavuuru	N-PRP-LOC
3	18-19	18-19	N-PRP-NUM
4	Karnataka	karnaaTaka	N-PRP-LOC
5	Tyagaraj	tyaagaraaja	N-PRP-PERSN
6	Syamashastri	syamasyastri	N-PRP-PERSN
7	Muttuswami	muttuswami	N-PRP-PERSN
8	Dixit	dikshit	N-PRP-PERSN

Figure 6. Recognition Rate



We carried out the experiments by varying the testing data , no of words changed as 2423, 4203, 6537, and observe that, our precision and recall are good, and having nearer values.

Table 6: Confusion Matrix

Corpus Size in Words	TP	FP	FN	TN	Precision	Recall	F-measure
2423	136	29	18	2269	82.42%	88%	85%
4203	230	50	29	3944	83.42%	89%	86%
6537	403	70	41	6093	85.14%	90.7%	87%

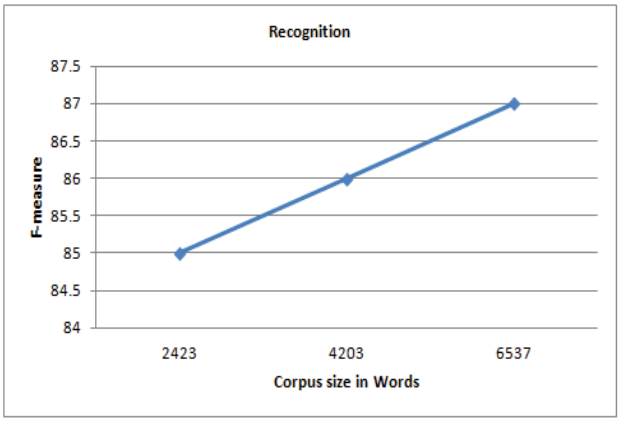


Figure 7. Corpus size versus F-mEasure

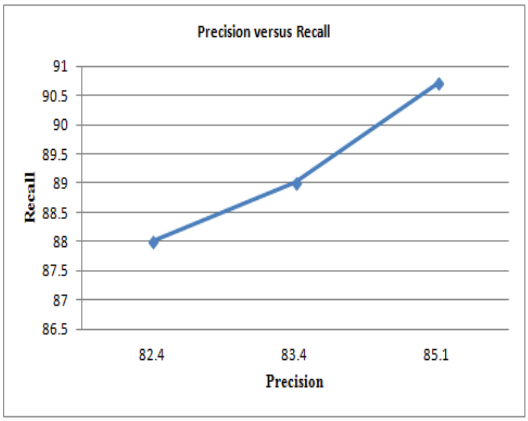


Figure 8. Precision versus Recall

VIII.CONCLUSION

Kannada poses challenge in NER due to its inherent ambiguity nature and lack of capitalization feature. The Kannada named entity recognition is a difficult task, especially to achieve human like performance. No work is cited in the literature on NER using rule based approach for Kannada. Our's is the first attempt, the proposed rule based methodology for recognition of Kannada named entities has good recognition rate and precision around 86%.. It is observed that use of suffix, prefix lists is important in identification of named entities. The performance can further be improved by improving gazetteer lists like proper noun dictionary, prefix and suffix.

REFERENCES

[1] Borthwick, "TA Maximum Entropy Approach to Named Entity Recognition". PhD thesis, NY University, September 1999.
[2] Michael Collins, "Ranking algorithms for named entity extraction: Boosting and the voted perception". Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp 489–496, 2002.
[3] Ekbal and S. Bandyopadhyay, "Named entity recognition in Bengali: A Conditional random field". Proc. ICON, pp 123–128, 2008.
[4] Michael Fleischman, "Automated sub categorization of named entities". Proc. Conference of the European Chapter of Association for Computational Linguistic, pp 25–30, 2001.

[5] Grishman, “The nyu system for muc-6”. Proc. Sixth Message Understanding Conference (MUC-6), pp 167–195, Fairfax, Virginia., 1995.
[6] Yungwei ding hsinhsi Chen and Shihchung TsaI, “Named entity extraction for information retrieval”. Proc. of HLT-NAACL,
[7] pp 8-15, 2003.
[8] Mukund Sangalikar, Shilpi Srivatsava and D.C. Kothari. ”Named entity recognition System for Hindi language”. International journal of Computational Linguistics Volume (2), pp 10–23.