Named Entity Recognition using Conditional Random Field for Kannada Language

Bhuvaneshwari C Melinamath

Abstract*— *Named Entity Recognition (NER) is a significant errand in Natural Language Processing (NLP) applications like Information Extraction, Question Answering and so on. In this paper, factual way to deal with perceive Kannada named substances like individual name, area name, association name, number, estimation and time is proposed. We have achieved higher accuracy in CRF approach than the in HMM approach. The accuracy of classification is more accurate in CRF approach due to flexibility of adding more features unlike joint probability alone as in HMM. In HMM it is not practical to represent multiple overlapping features and long term dependencies. CRF ++ Tool Kit is used for experimentation. The consequences of acknowledgment are empowering and the approach has the exactness around 86%.

Keywords: CRF, HMM, NER, NLP.

I. INTRODUCTION

Named Entity Recognition (NER) is an assignment of finding and arranging appropriate name, area name, association name, and so forth in a given text. Upper casing highlight is utilized in distinguishing named substances in English. But capitalization feature does not exist in Kannada and NER is an involved task in Kannada as compared to English. Proper nouns are indistinguishable from different forms of common nouns and adjectives in Kannada. These features of Kannada make NER a challenging task. The data extraction (IE) is crucial action in NER. On account of the significance of Information extraction (IE), DARPA (Defense Advanced Research Project Agency) started a number Massage Understanding Conferences (MUC) in the mid-nineties. As indicated by the particular characterized by MUC in (N. Chinchor, 1999), Machine learning methods include conditional random field, support vector machine (SVM), Maximum Entropy Model, Decision Tree, Hidden Markov Models (HMM) etc. In addition to machine learning techniques, one can use rule based approach, which needs effort in devising rules in consultation with experienced linguists. In machine learning procedures, we use accumulation of clarified records to prepare classifier. In any case, carefully assembled guideline based frameworks give great outcomes. In any case, the test lies being developed of guideline based strategies that require tremendous experience and linguistic learning of the language.

II. DESCRIPTION OF KANNADA LANGUAGE

Kannada language is a Dravidian language spoken essentially in Karnataka, a southern piece of India and positions third as far as number of speakers, as advised in registration data of 2013. Kannada is agglutinating language with complex morphology[5]. Morphology complexity of Kannada matches with Finnish and Turkish languages which are considered as complex morphology languages in. the world. Kannada language utilizes 49 phonemic letters, partitioned into 3 gatherings: 13 swaragaLu (called vowels in English), 34 vyaNjangaLu (called consonants in English) and 2 yogavaahakagaLu (neither consonants nor vowels), anusvara, to be specific, "aM " and visarga, in particular, "ah".

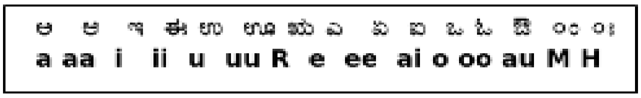
ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ ಅಂ ಅಃ
a aa i ii u uu R e ee ai o oo au M H

Figure No 1. Vowels in Kannada

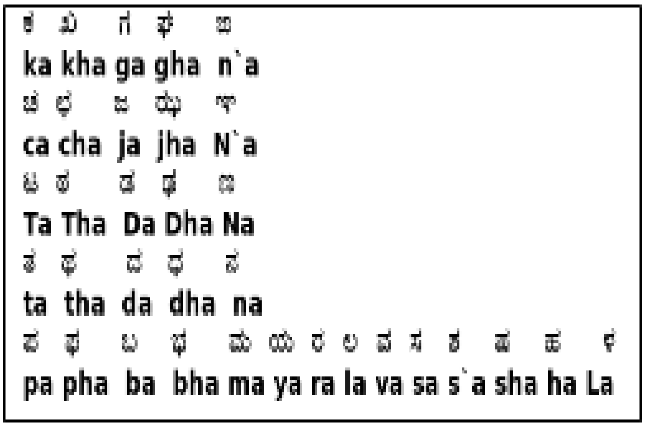
ಕ ಖ ಗ ಘ ಙ
ka kha ga gha n`a
ಚ ಛ ಜ ಝ ಞ
ca cha ja jha N`a
ಟ ಠ ಡ ಢ ಣ
Ta Tha Da Dha Na
ತ ಥ ದ ಧ ನ
ta tha da dha na
ಪ ಫ ಬ ಭ ಮ ಯ ರ ಲ ವ ಸ ಶ ಷ ಹ ಳ
pa pha ba bha ma ya ra la va sa s`a sha ha La

Figure No 2. Kannada Consonants

A solitary root word in Kannada can offer ascent to an enormous number of its surface structures [5]. The extravagance of Kannada lies in the way that noteworthy piece of sentence structure is taken care of by word morphology. For example, the transliterated form of Kannada word *'baravudillavenu'* is equivalent to several sentences / words in English, namely, ‘do you think he / she / they / it shall / will not come?. Therefore, words play an important role in Kannada language.

Revised Manuscript Received on August 05, 2019.

Bhuvaneshwari C Melinamath , Department of Computer Science & Engineering, SVERI’s, College of Engineering, Pandharpur, Maharashtra, INDIA. (Email: melinamathb@yahoo.com.)

Retrieval Number: K106609811S219/2019©BEIESP
DOI: 10.35940/ijitee.K1066.09811S219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

III. LITERATURE SURVEY

Named Entity Recognition (NER) has drawn more consideration of NLP scientists, since the most recent decade. Loads of work is done on NER for English utilizing the AI systems, both managed and solo. Following is the significance of works referred to in the writing. [2] has proposed a strategy for classification of area names utilizing Bayesian procedure and choice tree and the precision announced is 80%. [3] proposed a calculation for Named element for Information recovery. The review rates and the accuracy rates for the extraction of individual names, association names, and area names are (87.33%, 82.33%), (76.67%, 79.33%) and (77.00%, 82.00%), separately. [1] has proposed standard based calculation for Named Entity Recognition for Danish and reports F-measure as 92.7%. [6] have proposed a half and half approach, a blend of standard based, condition irregular field (CRF) and implied for named element acknowledgment for Hindi Language and reports accuracy of 96%,and review of 86.96%, f-proportion of 91%. [4] proposed a CRF based methodology for Bengali and reports F-score of 89.3 %. [8] has proposed a standard based methodology for Urdu utilizing little scale gazetteers and reports review of 90%, F-proportion of 93.14%, and exactness of 96.4%. [7] proposed guideline based methodology for Kannada. The aftereffects of acknowledgment are empowering and the technique has the exactness around 86%. From the literature survey, it is observed that not much work is carried out on NER for Indian languages, other than Hindi and Bengali. Little work is reported on NER for Kannada.

IV. PROPOSED METHODOLOGY

The proposed Conditional Random Field (CRF) NER system is shown in Figure 3.

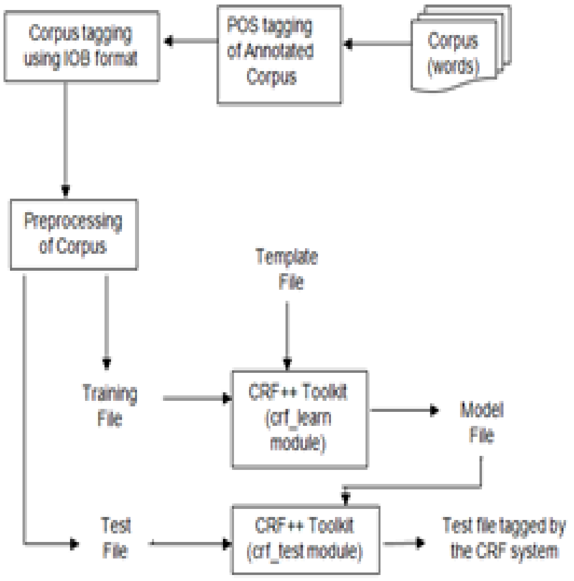
Corpus tagging using IOB format
POS tagging of Annotated Corpus
Corpus (words)
Preprocessing of Corpus
Template File
Training File
CRF++ Toolkit (crf_learn module)
Model File
Test File
CRF++ Toolkit (crf_test module)
Test file tagged by the CRF system

Figure No 3. Proposed CRF NER Architecture

The methodology takes a Kannada sentence as input, recognizes the named entities and categories them. The methodology comprises of tasks, namely, Corpus Tagging Scheme Using IOB (Inside outside Begin) Format, shown in table 1. Empty lines represent sentence boundaries. label B-XXX so as to demonstrate that it begins another element. The label I-XXX is utilized for words inside a named element of sort XXX. Words labeled with O are outside of named substances. Separate columns are attached for each list. Length feature is also added during this stage. After applying NE and POS tagging, the corpus is further preprocessed for adding Prefixes, Suffixes for each word. Gazetteer lists are also added here. If the word is present in any of the lists, then entry corresponding to that list is set to 1, otherwise it is 0.

Table No.1 IOB Tagging Scheme for CRF

Transliterated Tokens	POS	IOB
narendra	NNP	B-NEP
moodi	NNP	I-NEP
pradhaani	NN	NED
aadaru	VBZ	O

Another task is conversion of raw corpus into transliterated corpus using converter program (ir.pl), ir.pl is specially designed Perl program developed to convert Kannada Unicode text in to Roman form. This program transliterates the Kannada words based on phonemes.

CRF++ toolkit demands the data files to be in a particular format of multiple columns separated by single white space. Here the features are considered in 28 columns as shown in below box 1, and also in 3 column format. The primary section contains the present word, second word contains POS include in one segment, the third segment is IOB position which contains one segment, the fourth segment contains the word length and next segment contains the prefixes and additions. Prefix and addition window contains window size 5 to 7. The following is gazetteer list, which contains 12 sections. The last column is the true answer tag. The following are the feature samples considered here.

A. *Context Word.*

Parts of Speech (POS) Information.

B. *Word Prefix:*

We have experimented with a Word prefix, of length 1 to 4 characters, of the current word as a feature.

C. *Word Suffix:*

We have experimented with a Word suffix, of length 1 to 7 characters, of the current word as a feature.

D. *Named Entity Information:*

NE tag of the previous word and next word is used as a feature.

E. *Gazetteer Lists:*

We have developed the gazetteer lists.

- These rundowns have been utilized as the double esteemed highlights of the CRF.

- If the present token is in a specific rundown then the comparing highlight is set to 1 for the current and additionally the encompassing word(s), generally, set to 0.

Gazetteer list contain Location name, First name, Middle name, Last name, Organization name, Person Prefix, Day and Month lists, Abbreviation list etc. Consider the sentence below to show the sample format of training file.

F. English Text:

Yuvaraaj Singh is included in Asian Cup.

G. Transliteration:

yuuvaraaj siMganannu eshyaa kap Tiimnalli seerisikoLLalagide.

```
U00:%x[0,0]
U01:%x[-1,0]
U02:%x[1,0]
U03:%x[-2,0]
U04:%x[2,0]
U05:%x[0,1]
U06:%x[1,1]
U07:%x[-1,2]
```

CRF Toolkit makes the replacements in the template file during execution as per the token under consideration.

V. FEATURE ATTRIBUTE GENERATION

- w[t-2], w[t-1], w[t], w[t+1], w[t+2],
- w[t-1]|w[t], w[t]|w[t+1],
- pos[t-2], pos[t-1], pos[t], pos[t+1], pos[t+2],
- pos[t-2]|pos[t-1], pos[t-1]|pos[t], pos[t]|pos[t+1], pos[t+1]|pos[t+2],
- pos[t-2]|pos[t-1]|pos[t], pos[t-1]|pos[t]|pos[t+1], pos[t]|pos[t+1]|pos[t+2]

Here pos[t] and w[t] speak to the grammatical feature and word at position t in a grouping. The highlights considered express the normal for the word at position t by utilizing data from encompassing words, state w[t-1] and pos[t+1].

VI. RESULTS AND EXPERIMENTS

A sample file of size 1,000 words is considered, and it is observed that, the results obtained using CRF technique is more accurate as compared to HMM technique. But however the amount of time required in designing training and testing data is more in CRF technique. The technique works better if there is more training data. Here training data is tagged data, it is manually annotated corpus of 2,100 words. The number of named entities recognized is shown in table and figure 4. gives the number of entity types recognized in the input file.

Table No. 2 CRF Output of Named Entitles

Named Entity Types	No. of Named entities
NEP	48
NEL	14
NEO	25
NET	12
NEN	11
NED	25
NETE	3

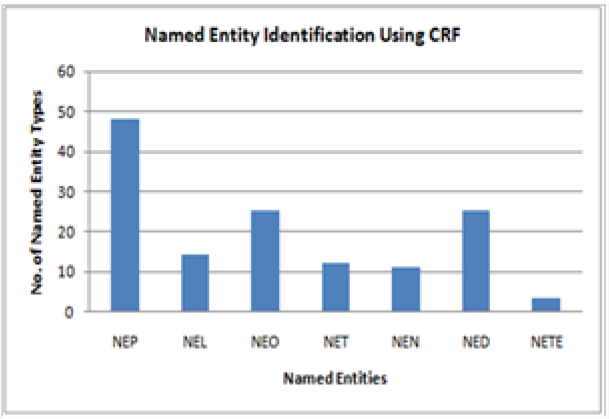

Figure No. 4. Named Entity Recognition by CRF

The named entity recognition for a file of 1,000 words is experimented by CRF method is shown in figure below. Accuracy of classification is more accurate in CRF in case of identifying organization entities as CRF is capable of identifying of long entities. However the size of organization entities varies, it is not of fixed length. The Results obtained are encouraging.

VII. ACKNOWLEDGMENT

I am very much thankful to our beloved Hon'ble Principal Dr. B. P. Ronge Sir, SVERI College of Engineering Pandharapur, for the continuous encouragement and support..

REFERENCES

1. Bick, Eckhard. "A Named Entity Recognizer for Danish". Proc. Conference on Language Resource and Evaluation 2004, pp. 305-308.
2. Michael Fleischman. "Automated sub categorization of named entities". Proc. Conference of the European Chapter of Association for Computational Linguistic, pp. 25-30, 2001.
3. Yungwei Ding Hsinhsi Chen and Shihchung tsai. "Named entity extraction for information retrieval". Proc. of HLT-NAACL, pp 8-15, 2003.
4. Ekbal and S. Bandyopadhyay ".Named entity recognition in Bengali: A Conditional random field". Proc. ICON, pp. 123-128, 2008.
5. Bhuvaneshwari C Melinamath (2011). "A robust Morphological analyzer to capture Kannada noun Morphology", Proc. IEEE, International Conference on Future Information Technology (ICFIT). Singapore, 2011.
6. 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, 2011.
7. Bhuvaneshwari C Melina math. "Rule Based Methodology for recognition of Kannada named entities," International journal of Latest trends in Engineering and technology (IJLTET). ISSN: 2278-621X. Vol. 3 Issue 4 March 2014.
8. Kashif Riaz, Proc. of " Named Entities Workshop", Uppsala, Sweden., pp. 126-135, Association for Computational Linguistics ACL, 16 July 2010 2010.

AUTHORS PROFILE

Bhuvaneshwari C Melinamath

Is working as HOD and professor in the Department of Computer Science and Engineering at SVERI COE Pandharpur, Maharashtra. She is having an experience of 25 years in teaching. She has completed Ph.D. from Central University of Hyderabad. She has contributed 5 lexical resources for Kannada. She has published around 30 papers in International Journals and conferences. She has received excellent paper award in ICFIT Conference held at Singapore in 2011. For the paper titled "A robust Morphological Analyzer for Kannada:" She has Received " Kittur Rani Cennamma Award " From Govt. of Karnataka from Chief Minister of Karnataka in 2017. She has also received "women of the year award in 2019 on International women Day celebration 8th March from Chief Minister of Karnataka. She is Member of IEI, ISTE, and IFREP. She has worked as member and president of many committees like BOS, DAB, and ICC etc.