Kannada Language Extraction Using Named Entity Recognition and Conditional Random Field

Fiza Kousar[1*], Chandrani Chakravorty[1]

[1]Department of Master of Computer Applications, RV College of Engineering, Bengaluru, Karnataka, India

Abstract

This paper deals with extraction of Kannada language data using Named Entity Recognition (NER) and Conditional Random Field (CRF) techniques. CRF methodology provided higher flexibility of adding new features to the Kannada language during the data extraction and increased the processing and response time when compared with Hidden Markov Models(HMM) of joint probability and support vector machine (SVM). The combination of the NER and CRF resulted in 86 % accuracy in data extraction. The Parts of Speech (POS) and Input Output Based (IOB) tagging were used to analyze and classify the words.

Keywords*: Named Entity Recognition, Conditional Random Fields, Kannada Language*

1.0 Introduction

Named entity recognition is used to identify and arrange the parameters such as person, location, association and so on in a given data. Named entity recognition is also majorly used for language translation. During the translation of the Kannada language, many language conflicts came into existence. While, the Kannada language has no segregation among the letters, English language is characterized with uppercase and lowercase letters. Also, the grammatical behavior of the letters can be easily differentiated in English whereas the syntactic functions of nouns and pronouns are indistinguishable from different forms of common nouns and adjectives in Kannada.

The extraction of Kannada letters in a grammatical manner from Named Entity Recognition is a challenging task which can be addressed by a suitable machine learning algorithms and methodologies[1]. Among the various algorithms and methodologies, the Conditional Random Field (CRF), Support Vector Machine (SVM) and Decision tree models provides accurate results for language data extraction. The supervised

[1]**Mail address: Fiza Kousar, MCA Student, Department of Master of Computer Applications, RV College of Engineering, Bengaluru – 59, Email: chandrani@rvce.edu.in, Ph:8904024234*

learning models of machine learning require the training data, based on which the classification of the data takes place. But, for the classification of any language data, linguistic knowledge is required for any model. Due to the flexibility of new features, the CRF can be trained can provide the boundary lines for the Kannada language to tag the key information which are extracted using Named Entity Recognition(NER). 'Tag' is the labeling of the elements to identify the parts of speech (POS). Each extracted key word is labeled as the transliterated token. The processed elements are labeled under two categories namely POS tag and IOB (Input Output Based) tag. POS tag is to identify the Parts of Speech of the processed element, whereas the IOB tag is used to classify whether the element is within a boundary or not. The boundary line is based on the element's syntactic function. For the languages such as Chinese and English, the boundary limit is not required as these languages are free to extract the key information without any linguistic knowledge. Based on the language and the grammatical manner, the CRF will adopt the new features. For Hindi language key data extraction, CRF can accept 12 tags[2]. This indicates the scalability and flexibility of the conditional random fields.

The Named entity recognition also takes advantage of Bidirectional long-short term memory which uses the character and sentence vector for the analysis [3]. The Bidirectional end to end sequence labeling is also used during the tagging for the extracted key word [4].The CRF also provides the flexibility of extracting the semantic elements from the Kannada language. This feature makes the CRF predominant over the other methodologies. Although, Hindi and Kannada languages have differences in pronunciation and grammatical word formation, the process to extract the data from both the languages is the same. Hence this paper focuses on implementation of conditional random field (CRF) methodology for the extraction of Kannada language data and categorize the words using Named Entity Recognition.

2.0 Kannada Language Description

Kannada language is majorly spoken in Karnataka state of south India. Kannada language has 49 phonemic letters which are higher than English alphabets [5]. These letters are broadly classified as 'Swara' (vowels), 'Vyanjana' (consonants), 'yogavahaka' (anuswaraAm and visarga Ah) as shown in Fig. 1 and 2.

Fig. 1. Kannada Vowels [6]

Fig. 2. Consonants in Kannada [7]

3.0 Data Extraction Methodology

The process of extraction of the keyword from the data is as shown in Fig. 3. For the Kannada language data extraction, the document file containing 15,000 data tests were adopted. Using these data tests, the system was trained and called as Data Training file. Once the data was collected from both the data test and the data training file, the system starts analyzing the data. This preprocessing and the analysis of the document was done using natural language processing. After the preprocessing and analysis, the key features were extracted from the data set and also from the training data using the Named Entity Recognition (NER). Once the features are extracted, each feature was labeled under supervised learning. Extracted key words from the larger data set were labeled with tags. Each tag represented the element's syntactic function. Based on the elements, grammatical behavior, labeling were performed. The syntactic functionality of the element was extracted as output and the same was evaluated based on the accuracy and exactness.

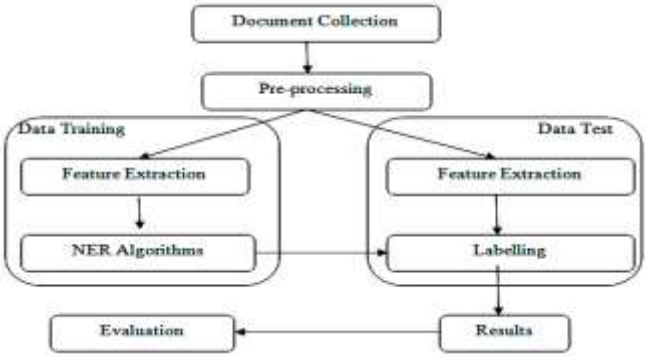

Fig. 3. Named Entity Recognition using Conditional Random FielddataExtraction[5]

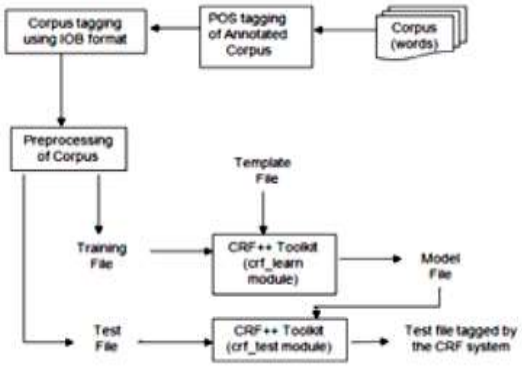

Fig. 4. CRF and NER Architecture[8]

The CRF and NER architecture as shown in Fig. 4 represents the process of the data extraction of any language using the training file which consists of the training data. The 'Corpus' refers to the words of the Kannada sentences among which the entities are recognized. This comprised of the task, training data file POS(Parts Of Speech) tagging and IOB (Inside outside Begin)format. POS tagging is used to identify the syntactic functional behavior of the pronoun and noun, which can be extracted using the NE(Noun Entity) Chunker. The IOB format was used to evaluate the corpus based on the syntactic functionality. Using the IOB format, Table 1 gives the extracted prefix and suffix for each word.

Processing of the Corpus(words) represents the sentence boundaries which are used to represent the empty lines. In order to identify the beginning of the new element, the labeling of the word must be needed. Every label of prefix represents the beginning of the word. Letter O will be used to categorize the element as the outside of the boundary. The boundary line is required for the Kannada language, which will vary from language to language. The elements which are not nouns and pronouns fall outside the boundary based on the training data file. In the training file, for each list, the separate columns will be represented. Those training files will be processed using the CRF++ Toolkit under the test modules method. The CRF++ toolkit was used to extract the data from the file. Once the file was uploaded into the CRF++ toolkit, using the Named Entity Recognition(NER) and Conditional Random Fields(CRF), the key elements were extracted and labeled using the POS and IOB tag. IOB tagging was done by analyzing and identifying the list of elements. The elements which are present within the list or boundary will be labeled. To identify whether the element is present in the list or not, the binary numbers are used. Digit 1 is to represent the element

within the list and digit 0 represents the element not outside of the list [5]. Based on this evaluation, the postfixes and suffixes were added to the list elements. Later, by using the Inside outside Begin format, the boundaries were identified and extracted.

Table 1. Input Output Based Tagging Scheme for CRF [8]

Transliterated Tokens	POS	IOB
Avalu	NN	NED
chennagi	NNP	I-NEP
Haadannu	NEP	B-NED
Haadidalu	VBZ	O
Adakkagi	VBZ	O
Avalige	NN	NED
Bahumanavannu	VBZ	O
Needidaru	VBZ	O

The Input Output Based scheme for the dataset uses the transliterated tokens as shown in Table 1. These transliterated tokens were the key words extracted using the named entity recognition. To label these key elements, a perl program was adopted for the Kannada data extraction. This program converts the Kannada unicode text into the Roman Form [5]. The empty lines along with labeled data are passed through the training file in which the first column represents the actual Kannada word, the second column represents the elements with added prefix and suffix and the third represents the Inside Outside Beginning words which means the boundary words. The data is collected as follows:

i) Prefix: The prefix was added to the word before 4 letters.
ii) Suffix: The suffix will be added within the 7 letters
iii) Named entity: The pre words and post words were tagged with the named entity treated as the feature

4.0 Analysis of Language Data

Some of the challenges were faced while extracting the data from Kannada language using NER and CRF. Eventhough some of the words like ಆದರು(Aadaru) and ಆದರೂ(Aadarū) are pronounced as same, the word ಆದರು(Aadaru) may not convey any meaning in the Kannada language and leads to error in the formation of statements. Hence, the grammatical behavior of the words also plays an important role during the Kannada language extraction. These minute changes are most important during the extraction for the accurate result. For this type of data extraction the CRF is best suited as it has the flexibility and scalability feature. In order to overcome the above challenges, more training data was adopted and every word was treated as the corpus. The count of the maximum number of corpus was used, which contains every

type of the words along with the smaller changes. In some of the statements such as ಅವಳುತಡವಾಗಿಬಂದಳು, the notation for the words were ಅವಳು/I-NEP ತಡವಾಗಿ/NN, NED, ಬಂದಳು/VBZ.O. Here, ಅವಳು is the pronoun, which is the femanine word, that is categorized as the proper pronoun as NEP and also boundaries were identified using CRF capabilities. The nouns and pronouns were classified within the boundary and other words were classified as out of the boundary. Based on the above classification, the sentence was categorized and extracted. The general comparison between CRF, HMM and SVM methodology is as shown in Table 2.

Table 2. Comparison between the CRF, HMM and SVM

Conditional Random Fields	Hidden Markov Models	Support Vector Machine
CRF approach has more accuracy due to the flexibility of adding more features unlike joint probability	HMM do not represent multiple overlapping features and long term dependency	SVM doesn't support for the large data set and lacks scalability feature
Conditional random field have the free flow data structure as, the boundary lines can be set for the required language and also can remove the boundary for the other languages	The HMM is dependent on every state and its corresponding observed objects. The flexibility feature is not applicable to HMM	Target classes are overlapped. Here the number of key features extracted are more than the training data set
The exactness for the kannada language extraction was 86%	The percentage for the generic named entity recognition was around 80.9%	The percentage of exactness for the Hindi and Bengali language was around 77.17%

5.0 Conclusion

This paper aimed at adoption of Named Entity Recognition (NER) for extracting the key element and Conditional Random field (CRF) for tagging the extracted element in Kannada Language. This combination of the NER and CRF makes the data extraction successful with around 86% of accuracy. The real need for the NER and CRF data extraction is the large training data which can help to analyze the words and categorize them along with the Parts Of Speech (POS) and Input Output Based (IOB) tagging.

References

01. B C Melinamath, Named Entity Recognition using Conditional Random Field for Kannada Language, *International Journal of Innovative Technology and Exploring Engineering(IJITEE),*8 (11S2), 413-416, 2019.

02. EAsif, B Sivaji, A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi, *Linguistic Issues in Language Technology,* 2 (1), 2009.

03. Nita Patil, Ajay Patil, B V Pawar, Named Entity Recognition using Conditional Random Fields, *Procedia Computer Science*, 167(6), 1181-1188, 2019.

04. X Ma, E Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, *In ACL*. 1, 1064–1074, 2016.

05. S Kale, S Govilkar, Survey of Named Entity Recognition Techniques for Various Indian Regional Languages, *International Journal of Computer Applications*, 164 (4), 0975 – 8887, 2017.

06. http://skoolshopblog.blogspot.com/2012/12/the-magic-of-kannada-language.html

07. https://asmeti.wordpress.com/learn-kannada/kannada-alphabets-numbers/

08. M L Patawar, M A Potey, Named Entity Recognition from Indian tweets using Conditional Random Fields based Approach, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5 (5), 1541-1545, 2016.