Statistical Dictionary with Conditional Random Fields to Identify the Kannada Named Entities

M. Pushpalatha, Antony Selvadoss Thanamani

Abstract--- We present an algorithm to recognize and identify the named entities of Kannada text document. The Kannada text document is collected from Central Institute of Indian Languages has many issues to be addressed. The proposed method has addressed the objective of algorithm is to determine and recognize the Kannada Named Entities like name of a person, designation of a person and place needs to be identified and recognized. The proposed statistical dictionary with conditional random fields in deep neural networks have been used to achieve the task of recognition of Kannada Named Entities The dictionary of Kannada words is formed from the statistical approach of matching patterns of Unicode values of individual words of a document. The sequence of Unicode values are considered for matching of patterns with deep architecture of neural networks has helped us in recognizing the Kannada word items from a dictionary formed from the proposed method. Finally the proposed method has achieved an accuracy of 84.46% from the proposed statistical dictionary of Kannada words with Conditional Random Fields.

Keywords--- CRF, Dictionary, Deep Learning, KNER.

I. INTRODUCTION

Natural Languages require processing if it is to be recognized by the machine learning system. Machine learning is a broad area of research, which mainly focuses on three different types of machine learning methodologies like supervised machine learning, Un-supervised machine learning, Semi-supervised machine learning.

All these methodologies have their own principles and steps to achieve the objectives of the research. Supervised machine learning [10], [19], [23], [24] is subject of interest of this research article, as we are focusing towards processing of natural languages like regional languages of India. Recognizing or identifying such entities of regional languages. While identifying such regional languages, we require a corpus of documents, which can be processed and identified with any of the machine learning methods.

Identification of Kannada sentences consisting of many grammatical words like noun, verb and object of a sentence is quite a challenging task, as the named entities may have been required to be identified with certain characters like votthakshara, dheerga, which is quite a challenging problem in Natural Language Processing [12], [13], [14], [15], [16].

Since NLP of certain languages like Kannada and other regional languages have a certain rules while forming a sentence unlike English. English has a simple rule of subject+ verb+object forms a sentence in English.

Similarly, there are certain specific rules, which we have to follow while forming a sentence in Kannada and other regional languages.

While forming these sentences, we need to recognize the words which a suitable as placeholder of a sentence.

The conditional random field is a method of machine learning approach [21], [19], [18], [17], which has been adopted to recognize and identify the phrases of Kannada language associated with entities.

The entities may be Name of a person, designation of a person, title of a person, place and other related entities. Such identification requires robust and efficient algorithms [20], [22], [12] to recognize the entities of Kannada languages.

The focus of our research article is to identify and recognize entities of Kannada regional language. Thus, we have adopted the method of CRF to meet the objective of the research.

The entire research article shall be visualized in various sections like section 2 presents the related work of kannada named entity recognition, section 3 provides a proposed method of identification of kannada named entities, section 4 discusses the results of the proposed method section 5 concludes the research article with few contribution towards kannada named entities.

II. RELATED WORK

The proposed research method has focused on aspects of Natural language processing based on deep neural networks in combination with conditional random fields. The research article [3], [5], [8] has focused on and gained attention of learning the conditional random fields using different classifiers, which has been used in our proposed method to enhance the features classification using a classifier support vector machine.

Further, research articles [4], [1], [6], [8], [9] has been studied and learned the extraction of features from documents, as to how the processing of documents are to be done before classification of any named entities.
Datasets

The Kannada Text document of CIIL used for processing

Proposed method worked with dataset of Central Institute of Indian Languages

Proposed Statistical Dictionary with Conditional Random Field

The dictionary of Kannada words are formed with the help of input, hidden and output layers of a machine learning methodology.

The proposed method involves different phases of computations, as it presents the objective of recognizing the Kannada Named Entities is done by tagging the words sequentially. The term sequentially indicates that the sorting of first character of a Kannada words are sorted sequentially based on Unicode values of each and every individual words of a text document. The text document may be comprised of different paragraphs, each paragraphs may consists of several characters. Such characters and words are identified with the help of a special character called spaces between words.

The notion of this research article can be very well presented, as to how the tagging can be done. Since, tagging is achieved by identifying full characters, half characters and special character spaces. Such characters are identified with the help of a special character called spaces between words. Such characters are identified with the help of a special character called spaces between words.

The dictionary of Kannada words are formed with the help of input, hidden and output layers of a machine learning methodology.

The proposed method involves different phases of computations, as it presents the objective of recognizing the Kannada Named Entities is done by tagging the words sequentially. The term sequentially indicates that the sorting of first character of a Kannada words are sorted sequentially based on Unicode values of each and every individual words of a text document. The text document may be comprised of different paragraphs, each paragraphs may consists of several characters. Such characters and words are identified with the help of a special character called spaces between words.

The notion of this research article can be very well presented, as to how the tagging can be done. Since, tagging is achieved by identifying full characters, half characters and special character spaces. Such characters are identified with the help of a special character called spaces between words. Such characters are identified with the help of a special character called spaces between words.
The concatenation of random probability distribution has been used in concatenation with deep learning to understand and analyze the words at different layers. The individual layers are trained with these items of dictionary in concatenation with weights of random probability distribution at every individual layers to achieve the desired objective of recognizing the words of a textual document.

The important point to be noticed is to understand, as to how the training is to be carried out at individual layers of a deep neural network.

**Training of deep neural network**

The deep neural network is trained; as such the individual layers are used to assign weights to annotated words of a sentence. The term individual layers represents that deep neural network can be assigned some weights based on the information gathered with preprocessing the textual document containing the sentences, words and characters. The weights are calculated based on the input vectors of a textual document.

![Calculation of weights based on input vectors of a Kannada textual document.](Image)

**Fig.3:** Calculation of weights based on input vectors of a Kannada textual document.

It is evident from the above fig.3 that the weights $w_1, w_2, \ldots w_n$ are calculated based on the input vectors represented by the words of sentences. The deep layer of a neural network is trained in concatenation with individual layers of a network. The input vectors processed with preprocessing phase makes the system analyze and understand, as to how the weights are to be calculated based on the information obtained from the input layers as a result of both weight and input vectors, the information is gathered from a dictionary of annotated words.

The training is carried at multiple hidden layers of calculated information, which is a dot product of weights and input layers, as we are required to refine the set of words into different entities like name of a person, place, and designation of a person. Such information of a textual document is gathered at different hidden layers. Instead of calculating and refining the hidden layers at single stage. We have incorporated the task of gathering the details of an annotated word into a specific identifier in multiple individual layers.

The multiple hidden layers are used to accurately extract and identify the annotated dictionary items into different classes. The classes include name of a persons, designation of a person, and places are three different classes of information are considered for classification of words in Kannada textual document. The information gathered from the hidden layers is given to the output layer, which predicts the words into different classes of words in a text document. The gathered information is obtained scores of values into different classes. The classification of words into different classes is done based on the scores obtained from the proposed method. The scores of the words with maximum values are used to classify the words into different classes. Class-1: Names of a person, Class-2: Designation of person, Class-3: Place are three different classes of words used in this proposed word.

**Testing**

It is another important phase of this proposed method, where the inputs are Kannada text documents given to the system and the output of this proposed method is a determination of words in documents into different classes as mentioned in the previous section. The testing of proposed method over a dataset TDIL obtained from the CIIL has been processed with different possibilities of testing versus training data items from a dictionary of Kannada words.

The process of testing versus training is carried out at a ratio of 5:95, where 5 percent of data is adopted for testing and the remaining 95% of data is used for training. Similarly 10:90 is another ratio of experimental protocol with 10 percent of data is used for testing and 90 percent of data is used for training. The experimental protocol is also used with a ratio of 15:85, 15% of data is for testing and 85% of data is for training of data. Another possible ratio of data used in our proposed method is 20:80, where 20 percent of data is used for testing and the remaining 80 percent of data is used for training.

<table>
<thead>
<tr>
<th>Table 1: Accuracies of proposed method versus other contemporary methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold</td>
</tr>
<tr>
<td>Fold 1</td>
</tr>
<tr>
<td>S M Ravi et al.</td>
</tr>
<tr>
<td>K P Pallavi et al.</td>
</tr>
</tbody>
</table>

The possible ratios of testing versus training are done to calculate the threshold value, where the accuracy of the proposed method is maximum. The possible set of options are considered in our proposed method to measure the accuracy of the proposed method over a dataset TDIL collected from the central institute of Indian languages. The documents collected from the CIIL have a varied set of Kannada words.

![CRF concatenated with deep neural features to classify the Kannada words into different classes of words.](Image)

**Fig.4:** CRF concatenated with deep neural features to classify the Kannada words into different classes of words.
It is clear from the above representation of fig.4. That phase 1, we performed the preprocessing, phase 2 embedding of feature vectors with CRF of neural networks. Phase 3 indicates the processing of computed information on multiple networks, phase 4. Indicates the output along with classified output of the proposed method.

III. RESULTS AND DISCUSSION

The proposed method has yielded an accuracy of 84.46% on a dataset TDIL collected from Central Institute of Indian Languages. Some of the challenges we faced while designing an algorithm and incorporating the proposed model into the system of different neural network layers. The deep neural networks has been concatenated with weights of the feature vectors has achieved a good results.

The fold 1 has yielded an accuracy of 81.3%, fold 2 with 83.2%, fold 3 with 83.4%, fold 4 with 88.1%, and fold 5 with an accuracy of 86.3%. Thus, a mean accuracy of 84.46% has been achieved on a dataset TDIL consisting of Kannada Text Documents. Each and every document has been processed with the proposed method of statistical approach. The folds of accuracies of proposed method against other contemporary methods are shown in table 1. The mean accuracy of the proposed method has a significance of 84.4% with other methods.

The proposed method shall be used and extended to work on other Kannada text documents to recognize and identify the named entities of documents. As we were intended to achieve the task of identifying the named entities in Kannada text document, we designed an algorithm with the help of conditional random fields along with statistical dictionary of Kannada words of a document.

The contribution of the proposed method shall be seen different ways:

1. The Kannada text documents collected from the CIIL shall be extended to other Kannada text documents as well.
2. Proposed method has been incorporated with deep neural features in concatenation with statistical dictionary of Kannada words.

The above contributions of the proposed method shall be considered as an important contribution towards achieving the objectives of the proposed method.

Ⅳ. CONCLUSION

The classification of Kannada words into different classes is based on the scores calculated from the hidden layers. The hidden layers are used with deep neural network to obtain the weights of the input vector of words. This is finally processed at multiple hidden layers and obtained an accuracy of 84.46% from the proposed method concatenation of input vectors with deep neural networks.

REFERENCES


