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.

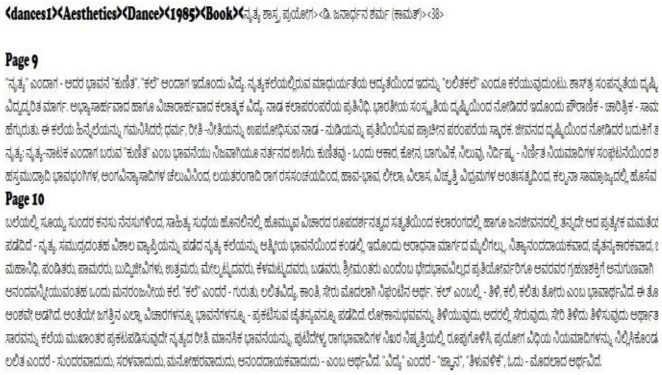
Further, research articles [4], [1], [6], [8], [9] has been studied and learned the extraction of features from documents, as to how the processing of documents are to be done before classification of any named entities.

Manuscript received February 01, 2019

M. Pushpalatha, Maharani's Science College for Women, Mysuru, Karnataka, India. (e-mail: pushpaharish78@gmail.com)

Dr. Antony Selvadoss Thanamani, HOD, Department of Computer Science, NGM College of Arts and Science, Pollachi, Bharathiar University, Coimbatore, Tamil Nadu, India. (e-mail: selvadoss@gmail.com)

Datasets



<dances1><Aesthetics><Dance><1985><Book><ನೃತ್ಯ ಶಾಸ್ತ್ರ ಪ್ರಯೋಗ><ಡಿ. ಜನಾರ್ಧನ ಶರ್ಮ (ಕಾಮತ್)><38>

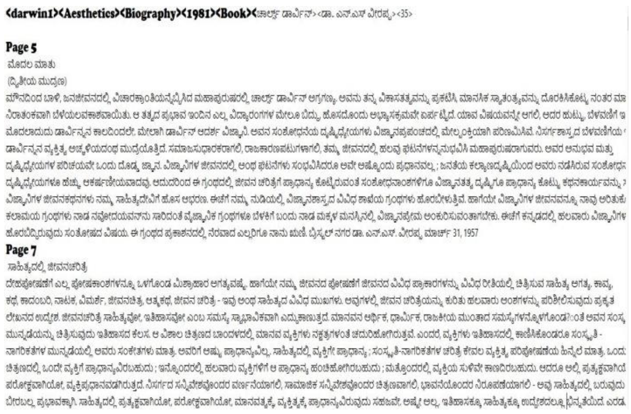
Page 9

"ನೃತ್ಯ" ಎಂದಾಗ - ಅದರ ಭಾವನೆ "ಕುಣಿತ". "ಕಲೆ" ಅಂದಾಗ ಇದೊಂದು ವಿದ್ಯೆ. ನೃತ್ಯಕಲೆಯಲ್ಲಿರುವ ಮಾಧುರ್ಯತೆಯ ಆದ್ಯತೆಯಿಂದ ಇದನ್ನು "ಲಲಿತಕಲೆ" ಎಂದೂ ಕರೆಯುವುದುಂಟು. ಶಾಸ್ತ್ರ ಸಂಪನ್ನತೆಯ ದೃಷ್ಟಿ
ವಿದ್ವದ್ಪುರಿತ ಮಾರ್ಗ. ಅಭ್ಯಾಸಾರ್ಹವಾದ ಹಾಗೂ ವಿಚಾರಾರ್ಹವಾದ ಕಲಾತ್ಮಕ ವಿದ್ಯೆ. ನಾಡ ಕಲಾಪರಂಪರೆಯ ಪ್ರತಿನಿಧಿ. ಭಾರತೀಯ ಸಂಸ್ಕೃತಿಯ ದೃಷ್ಟಿಯಿಂದ ನೋಡಿದರೆ ಇದೊಂದು ಪೌರಾಣಿಕ - ಚಾರಿತ್ರಿಕ - ಸಾಮ
ಹೆಗ್ಗುರುತು. ಈ ಕಲೆಯ ಹಿನ್ನೆಲೆಯನ್ನು ಗಮನಿಸಿದರೆ; ಧರ್ಮ, ರೀತಿ -ನೀತಿಯನ್ನು ಉಪಬೋಧಿಸುವ ನಾಡ - ನುಡಿಯನ್ನು ಪ್ರತಿಬಿಂಬಿಸುವ ಪ್ರಾಚೀನ ಪರಂಪರೆಯ ಸ್ಮಾರಕ. ಜೀವನದ ದೃಷ್ಟಿಯಿಂದ ನೋಡಿದರೆ ಬದುಕಿಗೆ ತ
ನೃತ್ಯ: ನೃತ್ಯ-ನಾಟಕ ಎಂದಾಗ ಬರುವ "ಕುಣಿತ" ಎಂಬ ಭಾವನೆಯು ನಿಜವಾಗಿಯೂ ನರ್ತನದ ಉಸಿರು. ಕುಣಿತವು - ಒಂದು ಆಕಾರ, ಕೋನ, ಬಾಗುವಿಕೆ, ನಿಲುವು, ನಿರ್ದಿಷ್ಟ - ನಿರ್ಣಿತ ನಿಯಮಾದಿಗಳ ಸಂಘಟನೆಯಿಂದ ಶ
ಹಸ್ತಮುದ್ರಾದಿ ಭಾವಭಂಗಿಗಳ, ಅಂಗವಿನ್ಯಾಸಾದಿಗಳ ಚೆಲುವಿನಿಂದ, ಲಯತರಂಗಾದಿ ರಾಗ ರಸಸಂಚಯದಿಂದ, ಹಾವ-ಭಾವ, ಲೀಲಾ, ವಿಲಾಸ, ವಿಚ್ಛಿತ್ತಿ ವಿಭ್ರಮಗಳ ಅಂತಃಸತ್ವದಿಂದ, ಕಲ್ಪನಾ ಸಾಮ್ರಾಜ್ಯದಲ್ಲಿ ಹೊಸವ

Page 10

ಬಲೆಯಲ್ಲಿ ಸೂಯ್ಯ ಸುಂದರ ಕನಸು ನೆನಸುಗಳಿಂದ, ಸಾಹಿತ್ಯ ಸುಧೆಯ ಹೊನಲಿನಲ್ಲಿ ಹೊಮ್ಮುವ ವಿಚಾರದ ರೂಪದರ್ಶನತ್ವದ ಸತ್ವತೆಯಿಂದ ಕಲಾರಂಗದಲ್ಲಿ ಹಾಗೂ ಜನಜೀವನದಲ್ಲಿ ತನ್ನದೇ ಆದ ಪ್ರತ್ಯೇಕ ಮಮತೆಯ
ಪಡೆದಿದೆ - ನೃತ್ಯ. ಸಮುದ್ರದಂತಹ ವಿಶಾಲ ವ್ಯಾಪ್ತಿಯನ್ನು ಪಡೆದ ನೃತ್ಯ ಕಲೆಯನ್ನು ಆತ್ಮೀಯ ಭಾವನೆಯಿಂದ ಕಂಡಲ್ಲಿ ಇದೊಂದು ಆರಾಧನಾ ಮಾರ್ಗದ ಮೈಲಿಗಲ್ಲು. ನಿತ್ಯಾನಂದದಾಯಕವಾದ, ಚೈತನ್ಯಕಾರಕವಾದ,
ಮಹಾನಿಧಿ. ಪಂಡಿತರು, ಪಾಮರರು, ಬುದ್ಧಿಜೀವಿಗಳು, ಉತ್ತಮರು, ಮೇಲ್ಪಟ್ಟವರು, ಕೆಳಮಟ್ಟದವರು, ಬಡವರು, ಶ್ರೀಮಂತರು ಎಂದೆಂಬ ಭೇದಭಾವವಿಲ್ಲದ ಪ್ರತಿಯೊರ್ವರಿಗೂ ಅವರವರ ಗ್ರಹಣಶಕ್ತಿಗೆ ಅನುಗುಣವಾಗಿ
ಆನಂದವನ್ನೀಯುವಂತಹ ಒಂದು ಮನರಂಜನೀಯ ಕಲೆ. "ಕಲೆ" ಎಂದರೆ - ಗುರುತು, ಲಲಿತವಿದ್ಯೆ, ಕಾಂತಿ, ಸೇರು ಮೊದಲಾಗಿ ನಿಘಂಟಿನ ಅರ್ಥ. 'ಕಲ್' ಎಂಬಲ್ಲಿ - ತಿಳಿ, ಕಲಿ, ಕಲಿತು ತೋರು ಎಂಬ ಭಾವಾರ್ಥವಿದೆ. ಈ ತೊ
ಅಂಶವೇ ಅಡಗಿದೆ. ಅಂತೆಯೇ ಜಗತ್ತಿನ ಎಲ್ಲಾ ವಿಚಾರಗಳನ್ನೂ ಭಾವನೆಗಳನ್ನೂ - ಪ್ರಕಟಿಸುವ ಚೈತನ್ಯವನ್ನೂ ಪಡೆದಿದೆ. ಲೋಕಾನುಭವವನ್ನು ತಿಳಿಯುವುದು, ಅದರಲ್ಲಿ ಸೇರುವುದು, ಸೇರಿ ತಿಳಿದು ತಿಳಿಸುವುದು ಅರ್ಥಾತ
ಸಾರವನ್ನು ಕಲೆಯ ಮುಖಾಂತರ ಪ್ರಕಟಪಡಿಸುವುದೇ ನೃತ್ಯದ ರೀತಿ. ಮಾನಸಿಕ ಭಾವನೆಯನ್ನು, ಪುಟಿದೇಳ್ವ ರಾಗಭಾವಾದಿಗಳ ನಿಖರ ನಿಷ್ಪತ್ತಿಯಲ್ಲಿ ರೂಪುಗೊಳಿಸಿ, ಪ್ರಯೋಗ ವಿಧಿಯ ನಿಯಮಾದಿಗಳನ್ನು ನಿಲ್ಲಿಸಿಕೊಂಡ
ಲಲಿತ ಎಂದರೆ - ಸುಂದರವಾದುದು, ಸರಳವಾದುದು, ಮನೋಹರವಾದುದು, ಆನಂದದಾಯಕವಾದುದು - ಎಂಬ ಅರ್ಥವಿದೆ. "ವಿದ್ಯೆ" ಎಂದರೆ - "ಜ್ಞಾನ", "ತಿಳುವಳಿಕೆ", ಓದು - ಮೊದಲಾದ ಅರ್ಥವಿದೆ.

Fig.1: Kannada Text document of CIIL used for processing



<darwin1><Aesthetics><Biography><1981><Book><ಚಾರ್ಲ್ಸ್ ಡಾರ್ವಿನ್><ಡಾ. ಎನ್.ಎಸ್ ವೀರಪ್ಪ><35>

Page 5

ಮೊದಲ ಮಾತು
(ದ್ವಿತೀಯ ಮುದ್ರಣ)
ಮೌನದಿಂದ ಬಾಳಿ, ಜನಜೀವನದಲ್ಲಿ ವಿಚಾರಕ್ರಾಂತಿಯನ್ನೆಬ್ಬಿಸಿದ ಮಹಾಪುರುಷರಲ್ಲಿ ಚಾರ್ಲ್ಸ್ ಡಾರ್ವಿನ್ ಅಗ್ರಗಣ್ಯ. ಅವನು ತನ್ನ ವಿಕಾಸತತ್ವವನ್ನು ಪ್ರಕಟಿಸಿ, ಮಾನಸಿಕ ಸ್ವಾತಂತ್ರ್ಯವನ್ನು ದೊರಕಿಸಿಕೊಟ್ಟ ನಂತರ ಮಾ
ನಿರಾತಂಕವಾಗಿ ಬೆಳೆಯಲವಕಾಶವಾಯಿತು. ಆ ತತ್ವದ ಪ್ರಭಾವ ಇಂದಿನ ಎಲ್ಲ ವಿದ್ಯಾರಂಗಗಳ ಮೇಲೂ ಬಿದ್ದು, ಹೊಸದೊಂದು ಅಭ್ಯಾಸಕ್ರಮವೇ ಏರ್ಪಟ್ಟಿದೆ. ಯಾವ ವಿಷಯವನ್ನೇ ಆಗಲಿ, ಅದರ ಹುಟ್ಟು, ಬೆಳವಣಿಗೆ ಇ
ಮೊದಲಾದುದು ಡಾರ್ವಿನ್ನನ ಕಾಲದಿಂದಲೇ. ಮೇಲಾಗಿ ಡಾರ್ವಿನ್ ಆದರ್ಶ ವಿಜ್ಞಾನಿ. ಅವನ ಸಂಶೋಧನೆಯ ದೃಷ್ಟಿಧ್ಯೇಯಗಳು ವಿಜ್ಞಾನಪ್ರಪಂಚದಲ್ಲಿ ಮೇಲ್ಪಂಕ್ತಿಯಾಗಿ ಪರಿಣಮಿಸಿವೆ. ನಿಸರ್ಗಶಾಸ್ತ್ರದ ಬೆಳವಣಿಗೆಯ
ಡಾರ್ವಿನ್ನನ ವ್ಯಕ್ತಿತ್ವ ಅಚ್ಚಳಿಯದಂಥ ಮುದ್ರೆಯೊತ್ತಿದೆ. ಸಮಾಜಸುಧಾರಕರಾಗಲಿ, ರಾಜಕಾರಣಪಟುಗಳಾಗಲಿ, ತಮ್ಮ ಜೀವನದಲ್ಲಿ ಹಲವು ಘಟನೆಗಳನ್ನನುಭವಿಸಿ ಮಹಾಪುರುಷರಾಗುವರು. ಅವರ ಅನುಭವ ಮತ್ತು
ದೃಷ್ಟಿಧ್ಯೇಯಗಳ ಪರಿಚಯವೇ ಒಂದು ದೊಡ್ಡ ಜ್ಞಾನ. ವಿಜ್ಞಾನಿಗಳ ಜೀವನದಲ್ಲಿ ಅಂಥ ಘಟನೆಗಳು ಸಂಭವಿಸಿದರೂ ಅವೇ ಅಷ್ಟೊಂದು ಪ್ರಧಾನವಲ್ಲ ; ಜನತೆಯ ಕಲ್ಯಾಣದೃಷ್ಟಿಯಿಂದ ಅವರು ನಡೆಸಿರುವ ಸಂಶೋಧ
ದೃಷ್ಟಿಧ್ಯೇಯಗಳೂ ಹೆಚ್ಚು ಆಕರ್ಷಣೀಯವಾದವು. ಆದುದರಿಂದ ಈ ಗ್ರಂಥದಲ್ಲಿ ಜೀವನ ಚರಿತ್ರೆಗೆ ಪ್ರಾಧಾನ್ಯ ಕೊಟ್ಟಿರುವಂತೆ ಸಂಶೋಧನಾಂಶಗಳಿಗೂ ವಿಜ್ಞಾನತತ್ವ ದೃಷ್ಟಿಗೂ ಪ್ರಾಧಾನ್ಯ ಕೊಟ್ಟು ಕಥನಕಾರ್ಯವನ್ನು
ವಿಜ್ಞಾನಿಗಳ ಜೀವನಕಥನಗಳು ನಮ್ಮ ಸಾಹಿತ್ಯದೇವಿಗೆ ಹೊಸ ಆಭರಣ. ಈಚೆಗೆ ನಮ್ಮ ನುಡಿಯಲ್ಲಿ ವಿಜ್ಞಾನಶಾಸ್ತ್ರದ ವಿವಿಧ ಶಾಖೆಯ ಗ್ರಂಥಗಳು ಹೊರಬೀಳುತ್ತಿವೆ. ಹಾಗೆಯೇ ವಿಜ್ಞಾನಿಗಳ ಜೀವನವನ್ನೂ ನಾವು ಅರಿತುಕೊ
ಕಲಾಮಯ ಗ್ರಂಥಗಳು ನಾಡ ನವೋದಯವನ್ನು ಸಾರಿದಂತೆ ವೈಜ್ಞಾನಿಕ ಗ್ರಂಥಗಳೂ ಬೆಳಕಿಗೆ ಬಂದು ನಾಡ ಮಕ್ಕಳ ಮನಸ್ಸಿನಲ್ಲಿ ವಿಜ್ಞಾನಪ್ರೇಮ ಅಂಕುರಿಸುವಂತಾಗಬೇಕು. ಈಚೆಗೆ ಕನ್ನಡದಲ್ಲಿ ಹಲವಾರು ವಿಜ್ಞಾನಿಗಳ
ಹೊರಬಿದ್ದಿರುವುದು ಸಂತೋಷದ ವಿಷಯ. ಈ ಗ್ರಂಥದ ಪ್ರಕಾಶನದಲ್ಲಿ ನೆರವಾದ ಎಲ್ಲರಿಗೂ ನಾನು ಋಣಿ. ಬ್ರಿಸ್ಟಲ್ ನಗರ ಡಾ. ಎನ್.ಎಸ್. ವೀರಪ್ಪ ಮಾರ್ಚ್ 31, 1957

Page 7

ಸಾಹಿತ್ಯದಲ್ಲಿ ಜೀವನಚರಿತ್ರೆ
ದೇಹಪೋಷಣೆಗೆ ಎಲ್ಲ ಪೋಷಕಾಂಶಗಳನ್ನೂ ಒಳಗೊಂಡ ಮಿಶ್ರಾಹಾರ ಅಗತ್ಯವಷ್ಟೆ. ಹಾಗೆಯೇ ನಮ್ಮ ಜೀವನದ ಪೋಷಣೆಗೆ ಜೀವನದ ವಿವಿಧ ಪ್ರಾಕಾರಗಳನ್ನು ವಿವಿಧ ರೀತಿಯಲ್ಲಿ ಚಿತ್ರಿಸುವ ಸಾಹಿತ್ಯ ಅಗತ್ಯ. ಕಾವ್ಯ,
ಕಥೆ, ಕಾದಂಬರಿ, ನಾಟಕ, ವಿಮರ್ಶೆ, ಜೀವನಚಿತ್ರ, ಆತ್ಮಕಥೆ, ಜೀವನ ಚರಿತ್ರೆ - ಇವು ಅಂಥ ಸಾಹಿತ್ಯದ ವಿವಿಧ ಮುಖಗಳು. ಅವುಗಳಲ್ಲಿ ಜೀವನ ಚರಿತ್ರೆಯನ್ನು ಕುರಿತು ಹಲವಾರು ಅಂಶಗಳನ್ನು ಪರಿಶೀಲಿಸುವುದು ಪ್ರಕೃತ
ಲೇಖನದ ಉದ್ದೇಶ. ಜೀವನಚರಿತ್ರೆ ಸಾಹಿತ್ಯವೋ, ಇತಿಹಾಸವೋ ಎಂಬ ಸಮಸ್ಯೆ ಸ್ವಾಭಾವಿಕವಾಗಿ ಎದ್ದುಕಾಣುತ್ತದೆ. ಮಾನವನ ಆರ್ಥಿಕ, ಧಾರ್ಮಿಕ, ರಾಜಕೀಯ ಮುಂತಾದ ಸಮಸ್ಯೆಗಳನ್ನೊಳಗೊಂಡ?ಂತೆ ಅವನ ಸಂಸ್ಕೃ
ಮುನ್ನಡೆಯನ್ನು ಚಿತ್ರಿಸುವುದು ಇತಿಹಾಸದ ಕೆಲಸ. ಆ ವಿಶಾಲ ಚಿತ್ರಣದ ಬಾಂದಳದಲ್ಲಿ ಮಾನವ ವ್ಯಕ್ತಿಗಳು ನಕ್ಷತ್ರಗಳಂತೆ ಚದುರಿಹೋಗಿರುತ್ತವೆ. ಎಂದರೆ, ವ್ಯಕ್ತಿಗಳು ಇತಿಹಾಸದಲ್ಲಿ ಕಾಣಿಸಿಕೊಂಡರೂ ಸಂಸ್ಕೃತಿ -
ನಾಗರಿಕತೆಗಳ ಮುನ್ನಡೆಯಲ್ಲಿ ಅವರು ಸಂಕೇತಗಳು ಮಾತ್ರ. ಅವರಿಗೆ ಅಷ್ಟು ಪ್ರಾಧಾನ್ಯವಿಲ್ಲ. ಸಾಹಿತ್ಯದಲ್ಲಿ ವ್ಯಕ್ತಿಗೇ ಪ್ರಾಧಾನ್ಯ ; ಸಂಸ್ಕೃತಿ-ನಾಗರಿಕತೆಗಳ ಚರಿತ್ರೆ ಕೇವಲ ವ್ಯಕ್ತಿತ್ವ ಪರಿಪೋಷಣೆಯ ಹಿನ್ನೆಲೆ ಮಾತ್ರ. ಒಂದು
ಚಿತ್ರಣದಲ್ಲಿ ಒಂದೇ ವ್ಯಕ್ತಿಗೆ ಪ್ರಾಧಾನ್ಯವಿರಬಹುದು ; ಇನ್ನೊಂದರಲ್ಲಿ ಹಲವಾರು ವ್ಯಕ್ತಿಗಳಿಗೆ ಆ ಪ್ರಾಧಾನ್ಯ ಹಂಚಿಹೋಗಿರಬಹುದು ; ಮತ್ತೊಂದರಲ್ಲಿ ವ್ಯಕ್ತಿಯ ಸುಳಿವೇ ಕಾಣದಿರಬಹುದು. ಆದರೂ ಅಲ್ಲಿ ಪ್ರತ್ಯಕ್ಷವಾಗಿಯೆ
ಪರೋಕ್ಷವಾಗಿಯೋ, ವ್ಯಕ್ತಿಪ್ರಧಾನವಡಗಿರುತ್ತದೆ. ನಿಸರ್ಗದ ಸನ್ನಿವೇಶವೊಂದರ ವರ್ಣನೆಯಾಗಲಿ, ಸಾಮಾಜಿಕ ಸನ್ನಿವೇಶವೊಂದರ ಚಿತ್ರಣವಾಗಲಿ, ಭಾವನೆಯೊಂದರ ನಿರೂಪಣೆಯಾಗಲಿ - ಅವು ಸಾಹಿತ್ಯದಲ್ಲಿ ಬರುವುದು
ಬೀರಬಲ್ಲ ಪ್ರಭಾವಕ್ಕಾಗಿ. ಸಾಹಿತ್ಯದಲ್ಲಿ ಪ್ರತ್ಯಕ್ಷವಾಗಿಯೋ, ಪರೋಕ್ಷವಾಗಿಯೋ, ಮಾನವತ್ವಕ್ಕೆ, ವ್ಯಕ್ತಿತ್ವಕ್ಕೆ ಪ್ರಾಧಾನ್ಯವಿರುವುದು ಸಹಜವೇ. ಅಷ್ಟೇ ಅಲ್ಲ, ಇತಿಹಾಸಕ್ಕೂ ಸಾಹಿತ್ಯಕ್ಕೂ ಉದ್ದೇಶದಲ್ಲೂ ಭಿನ್ನತೆಯಿದೆ. ಎರಡ

Fig.2: 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 words, and each word 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 [2], [7], [11] is a first phase of the objective of this research work; we need to preprocess all the characters of a Kannada text document. The Kannada text document may be comprised of several words; such words are separated by a special character spaces. The spaces has been indicated with a specific Unicode value, if such Unicode value occurs in a documents, while processing each and every characters of a text document, the system recognizes that the end of a word in a sentence has occurred and adds a word or a sequence of Unicode values into a dictionary. This not only makes a dictionary, but also with few processing forms tagging of words. The tagging of Kannada words are done by eliminating the repeating words from dictionary. The repeated words are identified by matching of sequence of patterns of Unicode values of one word corresponding to all other words of sentences.

The dictionaries of words are formed by extracting the words from a Kannada text document and sorting the Unicode values results in word added to dictionary.

Before adding any word into a dictionary, it has to be processed with a pattern match of Unicode values. One vs. all words of a text document. If there is any pattern match is identified, we retain only one copy of the word in a dictionary. This process is repeated until all unique words are added into the dictionary. This process completes the first phase of the proposed method.

Next, we perform features extraction, which consists of extraction of certain set of features like one full Kannada character, half Kannada characters and combination of full and half Kannada characters. Here the feature extraction plays a vital role in understanding or analyzing the words of a sentence, certain sentences of a paragraph. As we are most essentially required to find the relation between subject and object or predicate of a sentence, such predictions are made with the help of annotated words of a dictionary, where the words were added in the previous phase of the proposed method.

The features extraction involves different complexities like identifying full characters, half characters and combination of full and half characters together forms a meaningful word in a sentence. ರಾಜಕೀಯ, ಕೃಷ್ಣರಾಜ. The first character ಕೃ is a combination of two characters in Kannada such as ಕ and ರ. But the Unicode values of votthakshara and single characters are different ರ. Thus, we need to model to investigate and identify every individual character with a specific Unicode value even a half character. Such characters are identified with the help of a features extractor. The model features extractor has a capability of identifying votthakshara (ಲ್ಲಿ), dheerga (ಕೀ) words. Here the entities like Name of a person, designation of a person, place are identified and classified with the help of SVM Classifier.

The annotated words of a character are recognized based on the information of sequence of Unicode values of a dictionary items. These dictionary items are labeled with some rules as well. The process of extracting annotated words involves assigning weights to the words of a sentence based on the random probability distribution. The probabilities of weights are assigned based on the occurrence of data item in a dictionary. Such occurrences of patterns of Unicode values are used to extract the information of a word from a sentence.

The random probability distribution has been used in concatenation with deep learning to recognize the words as names, designation, and place in a sentence.

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.

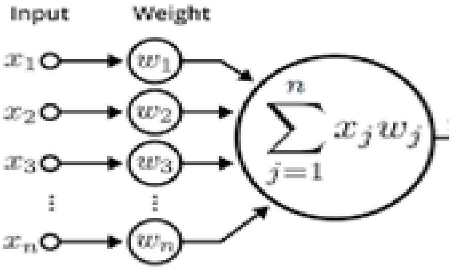


Fig.3: Calculation of weights based on input vectors of a Kannada textual document.

It is evident from the above fig.3 that the weights w_1, w_2…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 pre-processing 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.

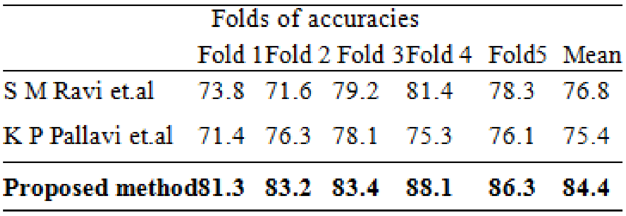
The multiple hidden layers are used to accurately extract and identify the annotated dictionary items into different classes. The classes include name of a persons, designation of a person, and places are three different classes of information are considered for classification of words in Kannada textual 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 1: Accuracies of proposed method versus other contemporary methods



	Folds of accuracies					
	Fold 1	Fold 2	Fold 3	Fold 4	Fold5	Mean
S M Ravi et.al	73.8	71.6	79.2	81.4	78.3	76.8
K P Pallavi et.al	71.4	76.3	78.1	75.3	76.1	75.4
Proposed method	**81.3**	**83.2**	**83.4**	**88.1**	**86.3**	**84.4**

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 TDL collected from the central institute of Indian languages. The documents collected from the CIIL have a varied set of Kannada words.

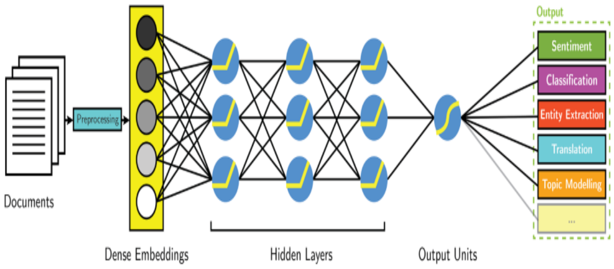


Fig.4: CRF concatenated with deep neural features to classify the Kannada words into different classes of words.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

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.

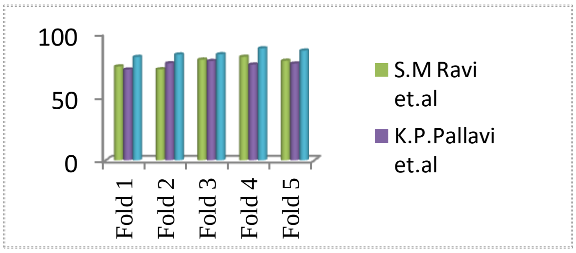


Fig.5: Accuracies of proposed method versus other contemporary methods

It is very clear from the above graphical representation of fig. 5 that the proposed method has shown good results in comparison with other approaches of Kannada Named Entity Recognition. Finally, we also state that the proposed method has yielded good accuracies in all the folds of data analysis.

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

1. Amarappa, S. and S.V. Sathyanarayana, 2013. Namedentity recognition and classification in Kannadalanguage. Int. J. Electron. Comput. Sci. Eng., 2:281-289.
2. Amarappa, S. and S.V. Sathyanarayana, 2013. A hybrid approach for Named Entity Recognition, Classification and Extraction (NERCE) in Kannada documents.Proc. Int. Conf. Multimedia Process.Commun.Info.Tech.
3. Amarappa, S. and S.V. Sathyanarayana, 2015. Kannada Named entity recognition and classification (nerc) based on multinomial naïve bayes (mnb) classifier. Int. J. Natural Language Comput. DOI: 10.5121/ijnlc.2015.4404
4. Bhat, S., 2012. Morpheme segmentation for Kannada standing on the shoulder of giants. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, (NLP' 12), pp: 79-94.
5. Bhuvaneshwari, C.M., 2014. Rule based methodology for recognition of kannada named entities. Int. J. Latest Trends Eng. Technol., 3: 50-59.
6. Cucerzan, S. and D. Yarowsky, 1999. Language independent named entity recognition combining morphological and contextual evidence. Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, (VLC' 99), pp: 90-99.
7. Curran, J.R. and S. Clark, 2003. Language independent NER using a maximum entropy tagger. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, (LLH' 03), Edmonton, pp: 164-167. DOI: 10.3115/1119176.1119200
8. Ekbal, A., R. Haque and S. Bandyopadhyay, 2008. Named entity recognition in Bengali: A conditional random field approach. IJCNLP.
9. Ekbal, A. and S. Bandyopadhyay, 2008. Bengali named entity recognition using support vector machine. IJCNLP.
10. Gali, K., H. Surana, A. Vaidya, P. Shishtla and D.M. Sharma, 2008. Aggregating machine learning and rulebased heuristics for named entity recognition. IJCNLP.
11. James, H., 1995. Natural Language Understanding. 1^{st}Edn., Dorling Kindersley pvt.Ltd., New Delhi, India.
12. Lafferty, J., A. McCallum and F.C. Pereira, 2001.Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
13. Malarkodi, C.S., R.K. Pattabhi and L.D. Sobha, 2012. Tamil NER - coping with real time challenges. Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages, (PIL' 12), pp: 23-23.

14. Murthy, V., M. Khapra and P. Bhattacharyya, 2016. Sharing network parameters for crosslingual named entity recognition. Comput.Sci. Nadeau, D. and S. Sekine, 2007.A survey of named entity recognition and classification. Lingvisticae Investigationes, 30: 3-26.
15. Nayan, A., B.R.K. Rao, P. Singh, S. Sanyal and R. Sanyal, 2008. Named entity recognition for Indian languages. Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages, (EAL' 08), pp: 97-104.
16. Noor, N.M., J. Sulaiman and S.A. Noah, 2016. Malay name entity recognition using limited resources. Adv. Sci. Lett., 22: 2968-2971. DOI: 10.1166/asl.2016.7124
17. Nothman, J., N. Ringland, W. Radford, T. Murphy and J.R. Curran, 2013. Learning multilingual named entity recognition from Wikipedia. Artificial Intell., 194: 151-175. DOI: 10.1016/j.artint.2012.03.006
18. Pallavi, K.P. and A.S. Pillai, 2015. Kannpos-kannada parts of speech tagger using conditional random fields. Proceedings of the Emerging Research in Computing, Information, Communication and Applications, (ICA' 15), Springer India, pp: 479-491.
19. Pandian, S., K.A. Pavithra and T. Geetha, 2007. Hybrid Three-stage named entity recognizer for Tamil. INFOS.
20. Pattabhi, R.K., T. Rao, S.R.R.R. Vijay, Vijayakrishna and L. Sobha, 2007. A text chunker and hybrid POS tagger for Indian languages. Shallow Parsing South Asian Languages. Riaz, K., 2010. Rule-based named entity recognition in Urdu. Proceedings of the 2010 Named Entities Workshop, Jul. 16-16, Uppsala, pp: 126-135.
21. Saha, S.K., S. Sarkar and P. Mitra, 2008a. A hybrid feature set based maximum entropy Hindi named Entity recognition. IJCNLP.
22. Saha, S.K., S. Chatterji, S. Dandapat, S. Sarkar and P. Mitra, 2008b. A hybrid approach for named entity Recognition in Indian languages. Proc. IJCNLP.
23. Shishtla, P., P. Pingali and V. Varma, 2008a. A character N-gram based approach for improved recall in Indian Language NER. IJCNLP.
24. Shishtla, P., K. Gali, P. Pingali and V. Varma, 2008b. Experiments in Telugu NER: A conditional random Field approach. IJCNLP, (2008, January) pp: 105-110.