Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021), 5731-5737

Research Article

Deep Learning Strategy to Recognize Kannada Named Entities

[1]M. Pushpalatha, [2]Dr. Antony Selvadoss Thanamani

[1]Maharani's Science College for Women (Autonomous), Mysuru, Karnataka, India.
(e-mail: pushpaharish78@gmail.com)
[2]HOD, Department of Computer Science, NGM College of Arts and Science, Pollachi, Bharathiar University, Coimbatore, Tamil Nadu, India.
(e-mail: selvadoss@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

ABSTRACT : Entity representatives are useful in understanding the natural language tasks including the semantics of the Kannada sentences into various entities. In this paper, we have come up with new pertained tag based representative learning of words and entities based on the bidirectional parsing. The proposed research works on segmenting the sentences of Kannada words into various taken, where every token makes various contributions in understanding the semantics of Kannada Sentences which treats words and entities in a given text as independent tokens, and outputs tagged entities based on representative learning mechanism. The research also has focused its attention towards achieving the results of good classification accuracy while recognizing the entities are through the tagging mechanism that is an extension of the general self-tagging mechanism of the Supervised Machine Learning Technique, and considers the types of tokens (words or entities) when computing attention scores. The erected research work has given its significant contribution in terms of good results over a standard benchmark datasets. In particular, it obtains state-of-the-art results on five well-known datasets: Open Entity (entity typing), TACRED (relation classification), CoNLL-2003 (named entity recognition), ReCoRD (cloze-style question answering), and SQuAD 1.1 (extractive question answering) as well as Kannada Named Entity Recognition of Central Institute of Indian Languages.

Keywords: Computer Science, Token Embedding, Position Embedding, Kannada Named Entity Recognition

1. INTRODUCTION

Natural Language processing tasks involves identification of Named Entities. These named entities may be identified in any languages such as Arabic, Persian, or any of the Indian Regional Languages. But these regional languages must be trained and annotated based on the semantics of the sentences in respective languages. The regional languages like Kannada, Hindim Gujarathi or any regional language has their own semantics of formation of sentences. [See fig.1] for more detailed information of processing of tagging and recognizing the named entities The applications of this research may be useful in different arenas such as commentary of the languages used in stadiums, questing and answer tagging, chat bot are certain critical and most essential applications that require the employability of this research Kannada Named Entity Recognition (KNER), The key aspect of this research includes identification of Kannada Named Entities are useful in benchmark dataset issued from Central Institute of Indian Languages (CIIL). This sector plays a significant role in training the teachers of various educational sectors. Hence, this research has provided its contributions in helping those training of teachers of Educational Institutions. Conventional entity representations assign each entity a fixed embedding vector that stores information regarding the entity in a knowledge base (KB) (Bordes et al., 2013; Trouillon et al., 2016; Yamada et al., 2016, 2017). This evolved research has thrown light on various aspects of Kannada named entities such as identification of Names of a person, Organization, Place and various other named entities of Kannada Sentences are recognized with the help of this Kannada Named Entity Recognition (KNER). Furthermore, the research has incorporated the ideas of various related areas of research arena that includes Knowledge base of different languages that needs to be processed. In contrast, to the contextual word representations (CWRs) defined by transformer (Vaswani et al., 2017), like BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2020), etc provide effective general-purpose word representations trained with unsupervised pretraining tasks based on language modeling.

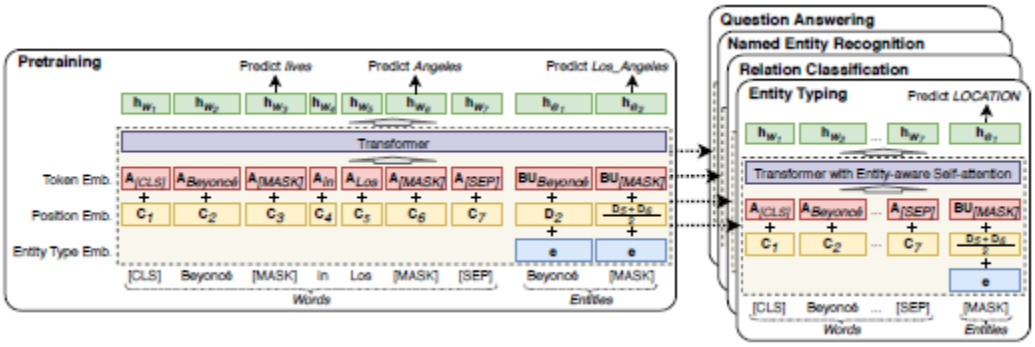

Figure 1: Architecture of LUKE using the input sentence “Beyonc´e lives in Los Angeles.” LUKE outputs contextualized representation for each word and entity in the text. The model is trained to predict randomly masked words (e.g., lives and Angeles in the figure) and entities (e.g., Los Angeles in the figure). Downstream tasks are solved using its output representations with linear classifiers.

The main contributions of this paper are summarized as follows:

• We propose DL-KNER, for recognizing the Kannada Named Entities. These Named Entities have been used for various benchmark applications like Understanding the Kannada Language, Language Question and Answer Tagging.

• We jave introduced a new strategy for training the systems to recognize the Kannada Named Entities using the Deep Learning Architecture. The SqueezeNet Architecture has been fine-tuned with a features descriptor before being processed with the system, which makes the system more efficient with better accuracy of results.

• DL-KNER has given its good empirical results in terms of accuracy and precision of recognizing the named entities on six popular datasets: Open Entity, TACRED, CoNLL- 2003, ReCoRD, and SQuAD 1.1 including the Kannada Named Entities (TDIL) provided from Central Institute of Indian Languages (CIIL).

2. DATASET

The research data consists of information collected from Central Institute of Indian Languages (CIIL). The CIIL has provided sufficient datasets in Kannada for the usefulness of training the staffs of Kannada Region, which is governed centrally by government of India. The research work has contributed the results on CIIL dataset for the purpose of assisting the staff of CIIL as a part of Natural Language Processing (NLP). Initially, the dataset is processed with certain operations like extraction of Kannada words from sentences. So that recognition of Kannada words and tagging is made possible.

<adlang4><Social Sciences><Linguistics><1984><Book><ಆಡಳಿತ ಭಾಷೆ ಕೆಲವು ವಿಚಾರಗಳು><ಪ್ರಧಾನ ಗುರುದತ್><46>

Page 104

(ಇ) ಕಾನೂನುಬಾಹಿರ ಸಂಸ್ಥೆಗಳು ಈಗ ಪ್ರಭಾವಹೀನವಾಗಿವೆ.
- ಎಂದು ಭಾಷಾಂತರಿಸಿಕೊಳ್ಳಬಹುದಾದರೂ, ಅವುಗಳಲ್ಲಿ ಸಹಜವಾಗಿ ಮೈದಳೆದ ಭಾಷೆಯ ಜೀವಂತಿಕೆ ಇಲ್ಲವೆಂಬುದು ಸ್ಪಷ್ಟವಾಗುತ್ತದೆ. ಇಂಥ ಸಂದರ್ಭಗಳಲ್ಲಿ (ಹಾಗೂ ಆದಷ್ಟು ಮಟ್ಟಿಗೂ ಎಲ್ಲ ಸಂದರ್ಭಗಳಲ್ಲಿಯೂ) ಕನ್ನಡದಲ್ಲಿಯೇ ಸ್ವತಂತ್ರವಾಗಿ ಚಿಂತಿಸಿ, ಕನ್ನಡ ನುಡಿಗಟ್ಟುಗಳನ್ನೇ ಬಳಸಿದರೆ ಅಂಥ ಭಾಷೆಯ ಸ್ವರೂಪ-ಶಕ್ತಿ-ಪ್ರಭಾವಗಳೇ ಬೇರೆಯಾಗುತ್ತವೆ. ಆರಂಭದ ಹಂತಗಳಲ್ಲೇನೋ, ಇಂಗ್ಲಿಷ್‌ನಲ್ಲಿ ಏನು ಬಳಸಿದೆ ಎಂಬುದನ್ನು ಗಮನಿಸಿ ಭಾಷಾಂತರಿಸಿಕೊಳ್ಳಬಹುದಾದರೂ, ಒಂದಲ್ಲ ಒಂದು ಹಂತದಲ್ಲಿ ಈ ಪ್ರಕ್ರಿಯೆಯನ್ನು ಕೈಬಿಟ್ಟು, ಸ್ವತಂತ್ರವಾಗಿ ಕಾರ್ಯ ನಿರ್ವಹಿಸತೊಡಗುವುದು ಅನಿವಾರ್ಯವಾಗುತ್ತದೆ. ಹಾಗೆ ಮಾಡುವುದು ಸಹಜವೂ ಹೌದು. ಈಗ ಮೇಲೆ ಹೇಳಿದ ವಾಕ್ಯಗಳಲ್ಲಿ, ಕನ್ನಡತನ ಹೇಗೆ ಮೈಗೂಡಿಕೊಳ್ಳಬಹುದೆಂಬುದನ್ನು ಗಮನಿಸೋಣ. ಆ ವಾಕ್ಯಗಳಲ್ಲಿ `beyond remedy', `to assault', `cease to be influential' ನಂಥ ನುಡಿಗಟ್ಟುಗಳು ಅಥವಾ ಪದಪುಂಜಗಳು ಇವೆ. ಅವುಗಳಿಗೆ ಸಂವಾದಿಯಾದ ನುಡಿಗಟ್ಟು ಅಥವಾ ಪದಪುಂಜಗಳನ್ನು ಕಲ್ಪಿಸಿಕೊಂಡರೆ ಆ ವಾಕ್ಯಗಳು ಹೀಗಾಗಬಹುದು:

(a)

(ಅ) ಪರಿಸ್ಥಿತಿ ಆಗಲೇ ಕೈಮೀರಿತ್ತು.
(ಆ) ಆ ಹೊತ್ತಿಗಾಗಲೇ ಪ್ರತಿಪಕ್ಷದವರು (ಅಥವಾ ಎದುರು ಗುಂಪಿನವರು)
ಅವರ ಮೇಲೆ ಕೈಮಾಡಿದ್ದರು.
(ಇ) ಈಗ ಕಾನೂನುಬಾಹಿರ ಸಂಸ್ಥೆಗಳ ಕೈನಡೆಯುವಂತಿಲ್ಲ.
ಆಡಳಿತ ಭಾಷೆಯನ್ನು ನಾವು ವಿಶ್ಲೇಷಿಸಬೇಕಾಗಿರುವುದು ಹಾಗೂ
ಅಳವಡಿಸಿಕೊಳ್ಳಬೇಕಾಗಿರುವುದು ಈ ದೃಷ್ಟಿಯಿಂದ ಎಂಬುದನ್ನು ಗಮನಿಸುವುದು ಒಳಿತು. ಏಕೆಂದರೆ
ಅಂಥ ವಿಶ್ಲೇಷಣೆ ಆಡಳಿತ ಭಾಷೆಯನ್ನು ನಾವು ಕನ್ನಡಕ್ಕೆ ಹೇಗೆ
ಕಸಿಮಾಡಿಕೊಳ್ಳಬಹುದು ಎಂಬುದನ್ನು ಮನದಟ್ಟು ಮಾಡಿಕೊಡುತ್ತದೆ; ನಮ್ಮ ಭಾಷೆಯ
ಶ್ರೀಮಂತಿಕೆಯನ್ನು ತೆರೆದು ತೋರುತ್ತದೆ. ಹೊರಭಾಷೆಗಳಿಂದ ಎರವಲು ಶಬ್ದಗಳಂತೆ
ಬಂದು, ಸಹಜ ಶಬ್ದಗಳೋ ಎಂಬಂತೆ ಕನ್ನಡದಲ್ಲಿ ಬೀಡುಬಿಟ್ಟಿರುವ
ಶಬ್ದಗಳು ಎಷ್ಟು ಶಕ್ತಿಪೂರ್ಣವಾಗಿ ಕಾರ್ಯ ನಿರ್ವಹಿಸುತ್ತವೆಂಬುದಕ್ಕೆ ಒಂದೆರಡು
ಉದಾಹರಣೆಗಳನ್ನು ಮಾತ್ರ ಇಲ್ಲಿ ಕೊಡಲಾಗಿದೆ:
(ಈ) ಅವನು ತನ್ನನ್ನು ಬಚಾವ್ ಮಾಡಿಕೊಳ್ಳಲಾರದೆ ಹೋದ
(ಉ) ಅವರ ಮೇಲೆ ನನಗೆ ಗುಮಾನಿ ಇತ್ತು

(b)

Page 105
ಈ ಸಂದರ್ಭಗಳಲ್ಲಿ `ಬಚಾವ್‌ಮಾಡು' (ಅಥವಾ `ಬಚಾಯಿಸು' ಎಂಬ ಇನ್ನೂ
ಕನ್ನಡೀಕರಿಸಿದ ರೂಪ), `ಗುಮಾನಿ' ಶಬ್ದಗಳಿಗೆ ಬದಲಾಗಿ `ರಕ್ಷಿಸಿಕೊಳ್ಳಲಾರದೆ
ಹೋದ', `ಸಂದೇಹವಿತ್ತು' ಎಂಬ ಶಬ್ದಗಳನ್ನು ಬಳಸಿದರೆ ಅನಾಹುತವೇನೂ
ಆಗುವುದಿಲ್ಲವೆಂಬುದು ನಿಜವಾದರೂ, ಈ ಶಬ್ದಗಳು ಆಮದಾದ ಆ ಶಬ್ದಗಳ ಎದುರಿನಲ್ಲಿ
ಪೇಲವವಾಗಿ ತೋರುತ್ತವೆಂಬ ಅಭಿಪ್ರಾಯವಂತೂ ಮೂಡಿಯೇ ಮೂಡುತ್ತದೆ.
ಸಮುಚ್ಚಯಗಳು: If, then, because, therefore, but, and
ಮೊದಲಾದ ಸಮುಚ್ಚಯಗಳನ್ನು (connectives) ವಾಕ್ಯದ ಆದಿಯಲ್ಲಿಯೇ ಬಳಸುವುದು
ಇಂಗ್ಲಿಷ್ ಭಾಷೆಯ ಜಾಯಮಾನ. ಕೆಲವು ಸಂದರ್ಭಗಳಲ್ಲಿ ಕೆಲವು ಸಮುಚ್ಚಯಗಳನ್ನು
ಆದಿಯಲ್ಲಿ ಬಳಸುವುದು ನಮ್ಮ ಭಾಷೆಗೂ ಹೊಂದಿಕೊಳ್ಳಬಹುದಾದರೂ, ಎಲ್ಲ
ಸಂದರ್ಭಗಳಲ್ಲಿಯೂ ಅವುಗಳನ್ನು ಯಾಂತ್ರಿಕವಾಗಿ ಭಾಷಾಂತರಿಸಿಕೊಂಡಲ್ಲಿ ಅವು
ಉದ್ದಿಷ್ಟ ಭಾಷೆಯ ಜಾಯಮಾನಕ್ಕೆ ಹೊಂದದ ಕೃತಕ ಪ್ರಯೋಗಗಳಾಗಿ ಬಿಡುತ್ತವೆ
ಎಂಬುದನ್ನು ಗಮನಿಸಬೇಕು. ಉದಾ:

(c)

Figure 2. Representation of dataset for Identification of Kannada Data with Tagging

The dataset shown in Figure 2 entails tagging of certain Kannada items based on the semantics of the sentence. The Kannada sentences are analysed by checking the grammar of respective sentences and predicts the appropriate meaningful tags to the learned information.

3. DEEP LEARNING FOR KANNADA NAMED ENTITY RECOGNITION (DL-KNER)

The research Deep Learning based Kannada Named Entities (DL-KNER) makes the system more efficient by replacing the conventional model of 3x3 filters by 1x1 point to point filters with fully connected neural strategies. These fully connected neural networks make the system more efficient by incorporating squeeze and expansion layers of Squeeze Net Architecture. The Network Architecture consists of Stack of fire module, which is more advantageous than AlexNet Architecture.

3.1. Description of Modified SqueezeNet for Kannada Named Entities

The research work has few significant contributions by replacing the conventional method of filtering. These elimination of filtering and incorporation of fire modules makes the system better for ascertaining the semantics of the Language.

• **Token embedding** presents the respective token with good understanding of analysis. The research has focused its attention by downsampling to keep the features of big size. The research has addressed the problem of analyzing and understanding the features of Kannada Tokens, every blank space makes a token in building the vocabulary of Kannada Sentences.

• **Position embedding** presents whether the position of the token is at the beginning or at the end of the sentence is notified with by analyzing the position of token. The subject position of Kannada sentence is same as predicate

position of English sentences. Similarly, the predicate position of Kannada Sentence is same as object part of English sentence. The figure.1 reflects the semantics of the language.

• **Entity embedding** the research work obtains the entities based on the information collected from the vocabulary constructed with Squeeze Net model. The fire module of the SqueezeNet Architecture makes the system more efficient while recognizing the Kannada Named Entity Recognition. The Kannada Named Entities has certain rules these rules are embedded to recognize the semantics of Kannada Sentences. [refer fig.1] for more detailed information of the evolved approach.

4. KANNADA NAMED ENTITY RECOGNITION (10 PT)

We have conducted certain experiments for measuring the accuracy of erected approach. The research has its significance in analysing the tokens as well as position of tokens in Kannada Sentences, which together makes the system more efficient in terms of precision and accuracy. The system is very well designed to analyse and understand the features of the tokens and positions of the tokens and finally with semantics of the language.

The fire module of the Squeeze Net architecture is very useful while analysing the tokens along with the positions of the tokens. The squeezing of features information along with expansion of features vectors makes the vocabulary better understandable. The vocabulary is built by following the semantics of the language mentioned in preceding section. The Kannada Named Entities has yielded good results with squeeze Net Architecture than any other techniques, as it involves certain parameters and rules to analyse and understand the features of the system.

The bunch of fire modules along with features representation makes the system more efficient. The fully connected neural strategy of Deep Learning is more advantageous than AlexNet Architecture, as it involves various fire modules consisting of squeeze and expansion of features. This squeeze and expansion of features are efficient for understanding the knowledge base.

The 1x1 filtering is more advantageous than 3x3 filtering, as it involves depth analysis for assessing the feature of the Kannada Sentences. The system is better than conventional approach, as it involves fully connected 1x1 filters as a bottleneck with depth of an information. Which is more convenient than other methods of convolutional approaches.

5. RESULTS AND DISCUSSIONS

This section provides a detailed information of results and their analysis while tagging the Kannada Named Entities. This also presents a graphical representation of research contributions in terms of measuring the accuracy of Kannada Named Entities (KNER). The Kannada Named Entities has its own impression the regions of India, where these language entities needs to be addressed with specific known semantics. These Semantics will help the system to recognize the Named words used in the semantically well written Kannada sentences.

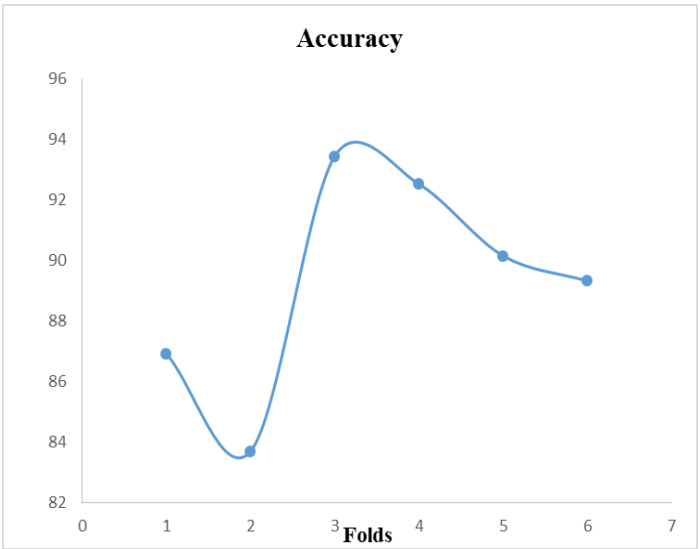

Figure 3. Accuracy of evolved method on TDIL dataset

We also conducted a research experiments on benchmark dataset TDIL provided by CIIL India for the purpose of ascertaining the efficacy of the research work. The overall performance of the research work has its impact on various other sectors of Kannada Named Entities. This has helped many while understanding the language. The significance of the research shall be seen in fig.3, fig.4 and fig.5. The respective results of analysis makes the system as state of the art by providing solutions to the various problems of Natural Language Processing especially in terms of identifying the Kannada Named Entities. These Kannada Named Entities involves identification of Terms such as name of the person, Organization of the person, Place, titles or awards of the person.
The Kannada Named Entities has its own impression the regions of India, where these language entities needs to be addressed with specific known semantics. These Semantics will help the system to recognize the Named words used in the semantically well written Kannada sentences. The overall performance of the research shall be seen in terms of an accuracy of 89.36% on standard benchmark dataset TDIL. The research contributions have been summarized in section 2 along with its performance.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

he eq. (1) , eq (2) and eq.(3) represents the performance metrics considered for evaluation of proposed method over a benchmark dataset TDIL provided by Central Institute of Indian Languages (CIIL).

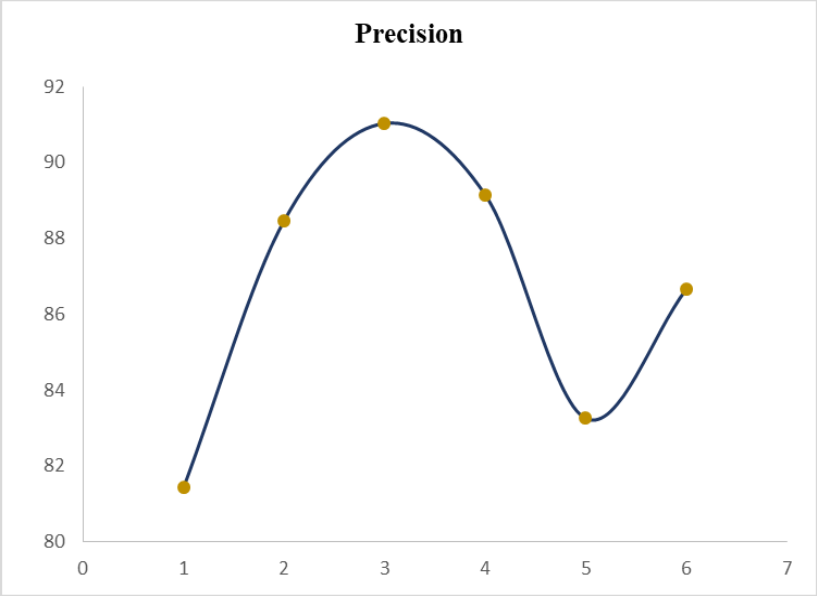

Figure 4. Precision of evolved method on TDIL dataset.

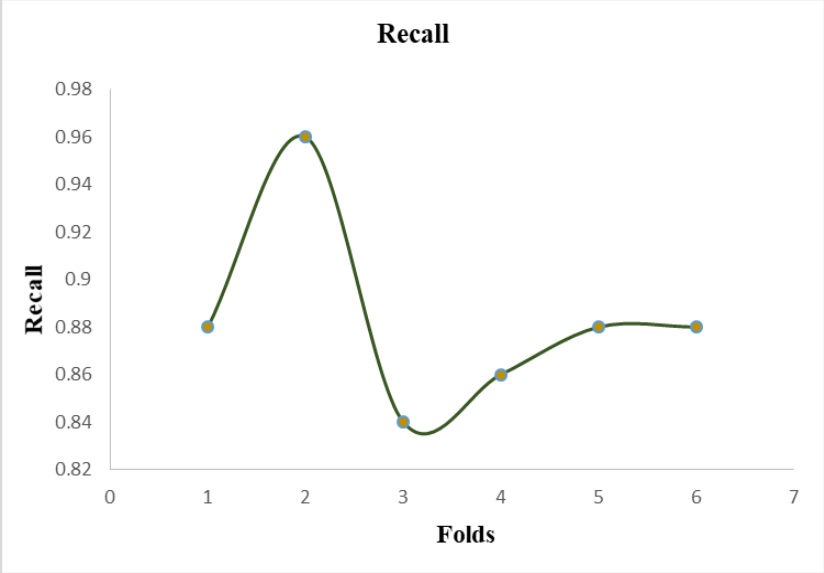

Figure 5. Recall of evolved method on TDIL dataset.

The results of accuracy along with precision and recall is shown here to represent the efficacy of the erected method of recognizing the Kannada Named Entities. [Refer fig.3] for accuracy of erected approach, [see fig. 4] for precision of the proposed method, [See fig.5] for recall of the erected method.

6. CONCLUSION

In this paper, we propose a Deep Learning based Kannada Named Entity Recognition (DL-KNER) based on benchmark dataset including TDIL of CIIL. The research has given its significant contributions in terms of good empirical results and performance measured from various metrics such as precision and recall in addition to accuracy of recognizing the Kannada Named Entities. The overall performance of the research shall be seen in terms of an accuracy of 89.36% on standard benchmark dataset TDIL. The research contributions have been summarized in section 2 along with its performance.

REFERENCES

1. X. S. Li, et al., "Analysis and Simplification of Three-Dimensional Space Vector PWM for Three-Phase Four-Leg Inverters," IEEE Transactions on Industrial Electronics, vol. 58, pp. 450-464, Feb 2011.
2. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet:
3. Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237v1.Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018a. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. arXiv preprintarXiv:1810.12885v1.
4. Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018b. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2205 – 2215.
5. Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Positionaware Attention and Supervised Data Improve Slot Filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 35–45.
6. Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1441–1451.
7. Xiepeng Li, Zhexi Zhang, Wei Zhu, Zheng Li, Yuan Ni, Peng Gao, Junchi Yan, and Guotong Xie. 2019. Pingan Smart Health and SJTU at COIN – Shared Task: utilizing Pre-trained Language Models and Common-sense Knowledge in Machine Reading Tasks. In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 93–98.
8. Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault F´evry, David Weiss, and Tom Kwiatkowski. 2020. Learning Cross- Context Entity Representations from Text. arXiv preprint arXiv:2001.03765v1.

9. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692v1.
10. Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark. 2019. Commonsense Inference in Natural Language Processing (COIN) - Shared Task Report. In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 66–74.
11. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In Proceedings of the 2018 Conferenceof the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227– 2237.
12. Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A.Smith. 2019. Knowledge Enhanced Contextual Word Representations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 43–54.
13. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to- Text Transformer. Journal of Machine Learning Research, 21(140):1–67.
14. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.
15. Emily Reif, Ann Yuan, MartinWattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim.2019. Visualizing and Measuring the Geometry of BERT. In Advances in Neural Information Processing Systems 32, pages 8594–8603.