ARTICLE

Cite This: J. of Artfic. Intell. and Machine. Lear. 2024, 2, 1-7

Sciforce

Journal of Artificial Intelligence and Machine Learning

Journal homepage: www.sciforce.org

Enhancing Named Entity Recognition in Low-Resource Dravidian Languages

Kiranmaye Panchadara

[1] *Kiranmaye Panchadara, Computer Science Engineer, Hyderabad, India*

ARTICLE INFO

Article history:
Received : 20241011
Received in revised form : 20241019
Accepted: 20241102
Available online : 20241112

Keywords:
Named Entity Recognition ;
Natural Language Processing, ;
Named Entity Recognition.

ABSTRACT

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP), yet its performance in low-resource languages, such as Kannada, Malayalam, Tamil, and Telugu of the Dravidian language family, remains challenging due to limited linguistic resources. Originating from India, these languages represent a rich linguistic diversity, but they often lack adequate resources for technological advancements. In this study, we explore methods to enhance NER performance in these low-resource Indian languages using multilingual learning and transfer learning techniques. Leveraging mBERT, RoBERTa, and XLM-RoBERTa algorithms, we conduct a comprehensive analysis. Initially, we evaluate each algorithm's performance on individual languages, obtaining accuracy scores. Subsequently, we merged datasets from pairs of languages to investigate cross-lingual transfer learning. For instance, combining Kannada and Tamil datasets yields a better accuracy, surpassing Kannada's standalone accuracy. We repeat this process for Tamil,Malayalam and Telugu subsequently, assessing both individual and multilingual accuracies. Our experiments provide insights into the efficacy of multilingual learning and transfer learning across diverse Dravidian languages, contributing to bridging the technological gap between urban and rural communities in India. By analyzing the impact of algorithm choice and cross-lingual transfer, we uncover valuable findings to advance NER performance in underrepresented languages. This study demonstrates the potential of technological advancements to empower diverse linguistic communities and foster inclusivity in NLP research and applications.

∗Corresponding author. e-mail: Kiranmaye2904@gmail.com

Introduction

Named Entity Recognition (NER) is a pivotal task in Natural Language Processing (NLP), indispensable for a spectrum of applications from information extraction to question answering systems. It entails identifying and classifying named entities, such as persons, locations, organizations, dates, and numerical expressions, within unstructured text. Accurate NER is crucial for a multitude of downstream tasks, including sentiment analysis, machine translation, and text summarization, contributing to enhanced understanding and interpretation of textual data. Advancements in artificial intelligence and the growing need for automated processing have underscored the significance of extracting semantic information from natural language texts 1. This task has become increasingly crucial as artificial intelligence endeavors to comprehend and interpret textual data more effectively.

Enhancing Named Entity Recognition in Low-Resource Dravidian Languages:

A Comparative Analysis of Multilingual Learning and Transfer Learning Techniques The named entity recognition model 2, built upon deep neural networks, has demonstrated promising outcomes across various named entity recognition tasks. However, its efficacy is contingent upon access to extensive training data annotated with tags.While significant strides have been made in NER for widely spoken languages such as English, Spanish, and Chinese, the scenario is starkly different for low-resource languages. Low-resource languages, characterized by limited linguistic resources and sparse annotated datasets, pose formidable challenges for NLP tasks, including NER. Indian languages have a limited amount of Named Entity Recognition (NER) literature. This scarcity primarily stems from data sparsity and a deficiency in tools due to the complexities

inherent in these languages. These complexities pose challenges for the adaptation of existing algorithms to low-resource languages 3.Among these low-resource languages, Dravidian languages stand out due to their linguistic diversity and historical significance. Kannada, Tamil, Telugu, and Malayalam are prominent Dravidian languages spoken by millions of people, primarily in Southern India and parts of Southeast Asia.

To address these challenges and mitigate the scarcity of data, researchers have turned to cross-lingual transfer training within the realm of named entity recognition. This approach focuses on preserving and transferring knowledge acquired while addressing one problem to another, albeit related, problem. It involves training a model on extensive corpora in one language and leveraging that knowledge to train a model on a smaller dataset in a different language. This strategy facilitates the improvement of entity classification in low-resource languages by leveraging data from multiple languages. Widely recognized as low-resource cross-lingual transfer learning 4, this method offers a promising avenue for advancing NER in diverse linguistic contexts. In this paper, we embark on a comprehensive exploration of NER in low-resource Dravidian languages, with a focus on Kannada, Tamil, Telugu, and Malayalam.

These four languages share significant linguistic similarities. While each language possesses its unique script and phonetic characteristics, they collectively exhibit common grammatical rules, vocabulary, and syntactic structures derived from their shared linguistic ancestry.Motivated by these linguistic affinities, our study undertakes a comprehensive cross-lingual evaluation to explore their potential for mutual assistance in Natural Language Processing (NLP) tasks, particularly in Named Entity Recognition (NER). Our research makes several noteworthy contributions:

We conduct an in-depth exploration of transfer learning architectures tailored for NER tasks, leveraging the linguistic similarities among Kannada, Tamil, Telugu, and Malayalam to enhance model performance.Through meticulous experimentation, we compare the efficacy of deep learning models trained using conventional monolingual approaches with those trained using multilingual strategies, elucidating the benefits of leveraging cross-lingual transfer learning across the Dravidian language family.We demonstrate the utility of cross-lingual transfer learning in improving NER outcomes for Kannada, Tamil, Telugu, and Malayalam, highlighting its role in harnessing linguistic similarities to bolster model performance and generalization across multiple languages.The structure of our paper unfolds as follows: Section 2 delves into the advancements in Named Entity Recognition, with a particular emphasis on the Dravidian language family and the linguistic similarities among Kannada, Tamil, Telugu, and Malayalam. In Section 3, we provide a detailed exposition of our experimental setup, elucidating the methodologies employed for evaluating different models in a cross-lingual context. The findings from our experiments are comprehensively summarized in Section 4, followed by concluding remarks in Section 5.

Related work

The concept of Named Entity Recognition was introduced by 5 at the Message Understanding Conferences (MUC) in the US. The exploration of this concept gained traction for Indian languages around 2008. Subsequent studies delved into employing various traditional machine learning algorithms and deep learning techniques. Notably, a comparison between Support Vector Machine (SVM) and Conditional Random Field (CRF) models, trained on identical datasets, revealed the superiority of the CRF model 6 Additionally, hybrid systems incorporating a blend of Hidden Markov Model, MaxEnt, and handcrafted rules demonstrated enhanced accuracy in NER tasks 7. As technology advanced, deep learning models became prevalent in addressing the NER problem. Models such as Convolutional Neural Network (CNN) 8The emergence of Cross-lingual Pre-trained Language Models marks a significant milestone in natural language processing, notably since the introduction of ELMO 9 and BERT 10. These models have garnered increasing attention due to their remarkable success in various NLP tasks, including but not limited to text classification, machine translation, text summarization, and question-answering systems. Their effectiveness has propelled them to the forefront of NLP research and application.

In tackling data sparsity, researchers have explored the utilization of cross-lingual resources 11, 12, 13 as a viable approach. Recently, multilingual learning using deep neural networks has emerged as a promising method for addressing data scarcity in low-resource languages 14. Multilingual Learning (MLL) can be viewed as a form of multi-task learning, wherein a deep neural network is trained for the same task across multiple languages by sharing some or all layers of the network among related languages. By leveraging shared layers (or weights) of the network, the model acquires cross-lingual features, enhancing its ability to generalize for the primary language task.

One of the notable advantages of multilingual learning with deep neural networks is its independence from parallel corpora, as it learns superior cross-lingual features from the training data of assisting languages. This approach is particularly effective when the languages involved are closely related.In the context of NER tasks in Indian languages, multilingual learning proves to be especially advantageous. Dravidian and Indo-Aryan languages in the Indian subcontinent exhibit shared phonemes, facilitating the establishment of correspondence between characters across scripts 15. They possess similar word order [15], enabling the sharing of syntactic and semantic information across languages. This characteristic allows for the sharing of sub-word information across languages in NER tasks. Moreover, Indian languages demonstrate high lexical overlap 16.[17] Introduces Naamapadam, the largest publicly available Named Entity Recognition (NER) dataset for the 11 major Indian languages from two language families. The dataset contains more than 400k sentences annotated with a total of at least 100k entities from three standard entity categories (Person, Locationand Organization) for 9 out of the 11 languages.[18]

addresses the challenge of data sparsity by leveraging characteristics from closely related language's data. The study employs hierarchical neural networks to train a supervised Named Entity Recognition (NER) system. Specifically, the network's shared layers are utilized to incorporate features from a closely related language. Multilingual Learning entails training a neural network on a merged dataset comprising a low-resource language and a closely related one. Our research aligns with theirs, as we amalgamate various datasets for Kannada, Malayalam, Tamil, and Telugu languages, subsequently comparing them against monolingual models.

Language	**O**	**PER**	**ORG**	**LOC**
Kannada	1037499	183414	87966	61201
Telugu	1050410	190212	104581	64430
Tamil	1373707	153773	122632	104457
Malayalam	902744	185849	77227	60966
Tamil with Kannada	3554688	578861	318895	237577
Tamil with Telugu	3556214	576325	320641	238478
Kannada with Tamil	3554688	578861	318895	237575
Telugu with Kannada	3267900	603864	291405	201663
Malayalam with Tamil	3534632	578833	316809	237785
Telugu with Tamil	3382799	601401	300058	202106

Table 1: TOTAL NUMBER OF INDIVIDUAL TAGS IN THE DATASETS

The datasets used for our study includes the Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages [17] datasets. To ensure consistency and comparability in our experiments, we streamlined the tagging process by considering only specific entity categories, namely location (LOC), person (PER), and organization (ORG). Additionally, we standardized the tagging format across datasets, converting I-PER, I-LOC, I-ORG, B-PER, B-LOC, B-ORG notations to PER,LOC and ORG respectively for uniformity. This involved replacing various tags with the 'O' tag for entities not of interest and mapping specific entity types to a common representation across datasets. We divided the data such that the split ensures that the majority of the data (70%) is allocated for training the model, while smaller portions are reserved for validating the model's performance during training (15%) and evaluating its generalization on unseen data (15%). In our research, we encountered several challenges while dealing with the datasets for Dravidian languages:Limited Annotated Data: One of the primary challenges was the scarcity of annotated data available for Dravidian languages, particularly for named entity recognition tasks. The existing datasets were relatively small and often insufficient to train robust models, leading to challenges in achieving high performance. Tagging was non-uniform, and the dataset included several confusing tags. In Also more than 33 percent of sentences contained tags.Cross-Lingual Differences: Dravidian languages exhibit significant linguistic diversity across regions and dialects, leading to challenges in building models that generalize well across different language varieties. Handling dialectal variations, morphological complexities, and lexical differences required careful consideration and adaptation of models.Our experimental setup involved merging and reshuffling datasets to create comprehensive training, tuning, and testing sets for evaluation. We employed transformer-based models such as mBert, Indicbert, and XLMRoBERTa to assess the performance of our NER models. By individually testing on specific test sets and conducting rigorous evaluations, we aimed to gauge the effectiveness of our approach in achieving accurate entity recognition and classification for Dravidian languages.

Experimental Setup

In our research focusing on Named Entity Recognition (NER), we concentrate on four prominent Dravidian languages: Kannada, Telugu, Malayalam, and Tamil, which are widely spoken in India. The NER task involves identifying and classifying named entities within text data, such as persons, organizations, and locations, which is crucial for numerous natural language processing applications. Leveraging publicly available datasets, we embarked on NER for Kannada, Telugu, Malayalam, and Tamil, aiming to explore the challenges and opportunities inherent in processing low-resource Dravidian languages.

In summary, our research underscores the significance of addressing the unique challenges posed by low-resource Dravidian languages in NER tasks. Through meticulous experimentation and careful data preprocessing, we endeavor to contribute valuable insights into thefeasibility and efficacy of NER for Kannada, Telugu, Malayalam, and Tamil, paving the way for enhanced natural language processing capabilities in these languages.

Methodology

In our study, we undertook a meticulous approach to dataset integration and model evaluation, aimed at maximizing the utility of available resources and ensuring robust performance assessment.Dataset Merging and Reshuffling: We merged the individual datasets, such as the Kannada and Tamil datasets, for each language under consideration. This consolidation process involved combining the training and tuning sets, followed by reshuffling to maintain randomness and prevent bias in subsequent analyses.Model Evaluation on Diverse Transformer-Based Architectures: Subsequently, we subjected the merged datasets to rigorous evaluation using a suite of transformer-based models, including mBERT, RoBERTa, and XLM-RoBERTa. Each model was individually trained on the integrated datasets and fine-tuned to capture language-specific nuances and entity recognition patterns.Individual Test Set Evaluation: To assess the models' generalization and performance, we conducted individual evaluations on distinct test sets associated with each dataset. For instance, models trained on the merged Kannada and Tamil datasets were tested individually on the respective test sets from the respective languages.

Cross-Lingual Evaluation: In addition to language-specific evaluations, we performed cross-lingual assessments by merging datasets from different languages, such as Kannada-Tamil, Tamil-Telugu, Tamil-Malayalam, Kannada-Telugu, Telugu-Tamil and Telugu-Kannada. This facilitated an examination of the models' adaptability across linguistic boundaries and their ability to transfer knowledge between related language pairs.Comprehensive Analysis: Finally, by evaluating the merged datasets individually on corresponding test sets, we gained insights into the models' proficiency in handling diverse linguistic structures and entity types.This systematic approach allowed us to derive meaningful conclusions regarding the effectiveness of transformer-based models for named entity recognition in low-resource languages.

Model Architecture

For our research paper on named entity recognition (NER) in low-resource Dravidian languages such as Kannada, Telugu, Malayalam, and Tamil, we can employ transformer-based architectures, which have demonstrated superior performance in NER tasks across various languages.Transformers have revolutionized the landscape of natural language processing, offering a powerful framework for processing sequential data such as text. These models learn representations of the data that capture its underlying structure, enabling them to excel in tasks like Named Entity Recognition (NER). In a transformer-based NER model, the transformer architecture plays a pivotal role in encoding input text into a sequence of hidden representations.

These representations are then fed into a classifier to predict named entities.One of the key features of transformer-based models is their ability to capture contextual relationships between words in the input text, achieved through a self-attention mechanism. This mechanism allows the transformer to attend to different parts of the input sequence and assign varying weights based on their relevance to the task at hand. Consequently, transformers can discern the context in which named entities appear, distinguishing whether they denote a person's name, an organization, or a location.Models like Multilingual BERT (mBERT), RoBERTa, and XLM-RoBERTa have further advanced transformer-based NER approaches. mBERT, trained on a diverse range of languages, offers cross-lingual capabilities, allowing it to generalize well across multiple languages. RoBERTa, an extension of BERT, employs improved pre-training techniques to enhance performance on downstream tasks like NER. XLM-RoBERTa builds upon RoBERTa by incorporating cross-lingual pre-training, enabling it to leverage multilingual data effectively.These transformer models are typically trained on large datasets of labeled examples, where they learn to identify the contextual cues surrounding named entities and associate them with the correct labels. They excel in handling sequence-to-sequence problems while also addressing long-range dependencies, thanks to their attention mechanism. By thoroughly examining inter-word connections within sentences, transformers can assign varying levels of significance to different components, facilitating swift and efficient resolution of ambiguities.

Multilingual BERT (mBERT): mBERT is a variant of the original BERT model trained on a diverse range of languages. mBERT is a variant of the BERT model trained on a diverse dataset comprising 104 languages.By leveraging data from multiple languages during training, mBERT can understand and process text in various languages simultaneously.Its multilingual capabilities enable mBERT to transfer knowledge across languages, making it effective for tasks involving multilingual data.mBERT's architecture allows it to capture cross-lingual similarities and differences, facilitating robust performance across diverse linguistic contexts.It aims to provide a single pre-trained model capable of performing well across multiple languages without the need for language-specific models.

Table 2: Accuracy Scores For The Models Tested On Both Individual And Megered Datasets

Language	**mBERT**	RoBERTa	XlmRoBERTa
Kannada	87.051	81.132	89.641
Telugu	92.889	80.459	95.139
Tamil	85.819	84.328	86.706
Malayalam	90.736	77.502	91.047
Tamil with Kannada	87.740	83.955	88.955
Tamil with Telugu	87.370	83.955	89.364
Kannada with Tamil	91.832	85.377	89.840
Telugu with Kannada	93.609	80.172	95.949
Malayalam with Tamil	90.947	78.142	91.884
Telugu with Tamil	94.149	79.855	94.959

By leveraging shared representations across languages during pre-training, mBERT demonstrates strong cross-lingual transfer capabilities.It has been widely used for various NLP tasks, including Named Entity Recognition (NER), machine translation, and sentiment analysis, among others.mBERT's ability to understand and generate text in multiple languages makes it a valuable resource for researchers and practitioners working with multilingual data.

RoBERTa: RoBERTa (Robustly optimized BERT approach) is an extension of the BERT model that incorporates several improvements in pre-training techniques. RoBERTa is an extension of the BERT model developed by Facebook AI.It is trained on a large corpus of English data using a self-supervised learning approach.RoBERTa improves upon BERT by introducing enhancements such as larger batch sizes, longer training duration, and removing the Next Sentence Prediction (NSP) task.By pre-training on raw texts without human labeling, RoBERTa learns rich contextual representations, leading to improved performance on downstream NLP tasks.It addresses limitations of BERT by using larger mini-batches, training for longer durations, and removing the next-sentence prediction task.RoBERTa achieves state-of-the-art performance on various NLP benchmarks by fine-tuning the pre-trained model on specific downstream tasks.With its enhanced pre-training methodology, RoBERTa captures richer contextual representations, leading to improved performance across a range of NLP tasks, including NER.RoBERTa has become a popular choice for researchers due to its robustness and effectiveness in handling diverse language tasks.

XLM-RoBERTa: XLM-RoBERTa (Cross-lingual Language Model - RoBERTa) is an extension of RoBERTa that incorporates cross-lingual pre-training.XLM-RoBERTa is a variant of the RoBERTa model developed by Facebook AI.It is pre-trained on a massive corpus of filtered CommonCrawl data

from 100 languages, making it a powerful multilingual language model.XLM-RoBERTa inherits the robustness and effectiveness of RoBERTa while extending its capabilities to multilingual contexts.By leveraging shared representations across languages, XLM-RoBERTa can effectively understand and generate text in multiple languages, making it valuable for cross-lingual NLP tasks.XLM-RoBERTa leverages shared representations across languages, allowing for effective transfer learning and generalization across diverse linguistic contexts.By pre-training on multilingual data, XLM-RoBERTa learns to capture universal linguistic patterns and representations, facilitating superior performance on cross-lingual tasks like NER.Its cross-lingual capabilities make XLM-RoBERTa particularly valuable for NLP applications involving multilingual data, where it outperforms monolingual models in terms of generalization and transfer learning.

Results

We now delve into our findings regarding the performance of multilingual models trained on the Dravidian languages (Tamil, Kannada, Telugu and Malayalam) NER datasets. Our objective is to assess the supportive capacity of these languages in enhancing the individual language performance.In this research endeavor, we aimed to evaluate the performance of named entity recognition (NER) models across four Dravidian languages: Tamil, Kannada, Telugu, and Malayalam. Initially, each language dataset underwent individual testing using three different algorithms: mBERT, XLM-RoBERTa, and RoBERTa. The assessments provided insights into the effectiveness of these algorithms across languages.Upon analyzing the results, it was observed that each language exhibited varying performance across the three algorithms. For instance, in Kannada, mBERT achieved relatively higher accuracy compared to XLM-RoBERTa and RoBERTa. Similarly, Tamil displayed a similar trend, with mBERT outperforming the other algorithms. In contrast, Telugu showcased superior performance with XLM-RoBERTa. In our comprehensive analysis of Named Entity Recognition (NER) across four Dravidian languages — Kannada, Tamil, Telugu, and Malayalam — using three distinct algorithms (mBERT, XLM-RoBERTa, and RoBERTa), several noteworthy observations emerged. Kannada demonstrated consistent high performance across all three algorithms, with mBERT yielding slightly better accuracy compared to XLM-RoBERTa and RoBERTa. Tamil, on the other hand, exhibited the highest accuracy among the languages, particularly with mBERT, while also displaying competitive performance with XLM-RoBERTa. Telugu demonstrated moderate accuracy overall, with XLM-RoBERTa outperforming mBERT and RoBERTa. Malayalam showcased impressive accuracy levels, particularly with mBERT and XLM-RoBERTa, outperforming RoBERTa. Subsequently, the datasets were merged to explore the impact of cross-lingual training on model performance. The merging process involved combining datasets from different languages and evaluating them using the same set of algorithms. Interestingly, the performance of merged datasets differed from individual language evaluations. When evaluated through cross-lingual transfer learning, merging Kannada with Tamil and Telugu resulted in improved performance across all algorithms, with XLM-RoBERTa showing notable enhancement. Tamil's merger with Telugu and Malayalam also demonstrated improved accuracy, particularly with RoBERTa. Despite discrepancies in dataset sizes, with Kannada and Malayalam datasets being comparatively smaller, merging datasets facilitated enhanced training and fine-tuning, contributing to improved performance metrics across all languages. This indicates the potential benefits of cross-lingual training in enhancing model performance.Furthermore, it is noteworthy that Kannada and Malayalam datasets were relatively smaller compared to Telugu and Tamil. Despite this data discrepancy, the merging of datasets allowed for enhanced training and fine-tuning, leading to improved performance metrics across languages.

Overall, these experiments shed light on the effectiveness of multilingual NER models and the advantages of cross-lingual training. The findings underscore the importance of considering language-specific nuances and dataset characteristics when developing NER models for low-resource languages. Additionally, the results provide valuable insights into optimizing model performance and lay the groundwork for future research in this area.

Conclusion

In conclusion, our research delved into the nuanced landscape of Named Entity Recognition (NER) across four Dravidian languages — Kannada, Tamil, Telugu, and Malayalam — leveraging three distinct algorithms: mBERT, XLM-RoBERTa, and RoBERTa. Through a meticulous evaluation process, we observed varying performance levels across languages and algorithms, with certain languages demonstrating consistent high accuracy and others exhibiting moderate to impressive results. We also investigated the collaborative potential of these languages, demonstrating that combining two languages yields improvements in individual language performance. Cross-lingual transfer learning further elucidated the potential for enhancing NER performance when merging datasets from different languages. The findings underscore the importance of considering language-specific nuances and dataset characteristics in NER model development and evaluation.

Looking ahead, future research avenues could explore the development of language-specific NER models optimized for Dravidian languages, potentially incorporating domain-specific knowledge and fine-tuning techniques to further enhance performance. Additionally, investigations into the transferability of pre-trained models across languages and domains could yield insights into broader applicability and generalization capabilities. Moreover, expanding the scope to include additional Dravidian languages and exploring multi-task learning frameworks could provide a more comprehensive understanding of NER challenges and solutions in the context of diverse linguistic landscapes. Ultimately, our research sets the stage for continued exploration and innovation in NER for

Dravidian languages, contributing to advancements in natural language processing and fostering broader language inclusivity in AI applications.

Funding Declaration:

Funding:
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. W **A. Bharadwaj, D. Mortensen, C. Dyer, J. Carbonell.** "Phonologically Aware Neural Model for Named Entity Recognition in Low Resource Transfer Settings." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, November 1–5, 2016. Pages 1462–1472.
2. **C.T. Tsai, S. Mayhew, D. Roth.** "Cross-Lingual Named Entity Recognition via Wikification." In *Proceedings of the Conference on Computational Natural Language Learning*, Berlin, Germany, January 1, 2016. Pages 219–228.
3. **R. Joshi.** "L3cube-mahanlp: Marathi natural language processing datasets, models, and library." *arXiv preprint* arXiv:2205.14728, 2022.
4. **S. Schuster, S. Gupta, R. Shah, M. Lewis.** "Cross-lingual transfer learning for multilingual taskoriented dialog." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019. Pages 3795–3805. [Online]. Available: https://aclanthology.org/N19-1380
5. **R. Grishman, B. Sundheim.** "Message Understanding Conference6: A brief history." In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. [Online]. Available: https://aclanthology.org/C96-1079
6. **A. A. Krishnarao, H. Gahlot, A. Srinet, D. S. Kushwaha.** "A comparative study of named entity recognition for Hindi using sequential learning algorithms." In *2009 IEEE International Advance Computing Conference*, 2009. Pages 1164–1169.
7. **R. Srihari.** "A hybrid approach for named entity and sub-type tagging." In *Sixth Applied Natural Language Processing Conference*, Seattle, Washington, USA, April 2000. Pages 247–254. [Online]. Available: https://aclanthology.org/A00-1034
8. **S. Albawi, T. A. Mohammed, S. Al-Zawi.** "Understanding of a convolutional neural network." In *2017 International Conference on Engineering and Technology (ICET)*, 2017. Pages 1–6.
9. **M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer.** "Deep contextualized word representations." *arXiv*, 2018. arXiv:1802.05365.
10. **J. Devlin, M.W. Chang, K. Lee, K. Toutanova.** "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv*, 2018. arXiv:1810.04805.
11. **Y. Chen, C. Zong, K.-Y. Su.** "A Joint Model to Identify and Align Bilingual Named Entities." *Computational Linguistics*, 39(2), June 2013. Pages 229–266. DOI: https://doi.org/10.1162/COLI_a_00122
12. **M. Faruqui.** ""Translation can't change a name": Using Multilingual Data for Named Entity Recognition." *arXiv preprint* abs/1405.0701, 2014. arXiv:1405.0701. http://arxiv.org/abs/1405.0701
13. **Z. Yang, R. Salakhutdinov, W. Cohen.** "Multi-Task Cross-Lingual Sequence Tagging from Scratch." *ICLR*, 2017.
14. **M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean.** "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." *Transactions of the Association for Computational Linguistics*, 2017.
15. **K.V. Subbārāo.** "South Asian Languages: A Syntactic Typology." 2012.
16. **A. Kunchukuttan, P. Bhattacharyya.** "Orthographic Syllable as basic unit for SMT between Related Languages." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, Texas, USA, November 1-4, 2016.
17. **A. Mhaske, H. Kedia, S. Doddapaneni, M.M. Khapra, P. Kumar, R. Murthy, A. Kunchukuttan.** "Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages." Indian Institute of Technology Madras, AI4Bharat, Microsoft India, IBM Research India.
18. **R. Murthy, M.M. Khapra, P. Bhattacharyya.** "Improving NER Tagging Performance in Low-Resource Languages via Multilingual Learning." *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 2, December 2018. [Online]. Available: https://doi.org/10.1145/3238797