# NER for Sanskrit Language

23M0837,Ujjwal Sharma, M.Tech 1st year,CSE

23M0753,Swapnil Bhattacharyya,Mtech 1st year, CSE

23M0834, Priyanshu Sharma, M.Tech 1st year, CSE

23D0373,Priya Mishra,Phd 1$^{st}$ year,CSE

# Problem Statement

- **Identifying Named Entities in Sanskrit Sentences**

Input: Sanskrit Sentence

Output: Named Entity Tagged Sentence

Example:

Input: श्रीकृष्णः द्वारिकायाः राजा अस्ति

Output: श्रीकृष्णः_(Person) द्वारिकायाः_(Location) राजा अस्ति

# Problem Statement | Sub Problems

- **Creating a dataset for Sanskrit NER**

- **Training  models for Sanskrit NER**

# Motivation of the problem

- There is no public NER dataset for Sanskrit.
- Creation of Dataset through human annotation is costly and time consuming.
- NER for Indian Languages has been limited to popular languages such as Hindi (***HiNER***; Murthy et al., 2022)

# Literature Survey

- IJCNLP 2008 NER Dataset: This dataset contains NER data in five languages, including Hindi. It has been used extensively in previous Hindi NER research.
- FIRE 2014 Dataset: Another dataset featuring NER data in four languages, including Hindi. It has been a resource for NER research.
- WikiANN Data: This dataset contains NER data in multiple languages, including Hindi. However, it is tagged automatically and considered a 'silver-standard' dataset for NER.
- Multilingual Transfer Learning: Rahimi et al. (2019) discuss the use of transfer learning for multilingual NER, evaluating results across multiple languages.
- Code-Mixed Hindi-English Text: Singh et al. (2018) employ Long Short-Term Memory (LSTM), Decision Trees, and Conditional Random Fields (CRF) for NER on code-mixed Hindi-English social media text.
- Hybrid Approaches: Past research has explored hybrid approaches combining CRF, Maximum Entropy (MaxEnt), and rules for Hindi NER using the IJCNLP-08 dataset.

# Literature Survey(Contd.)

- Contextualized Word Representations: Recent work includes the use of deep learning-based approaches. Singh et al. (2021) employ a Bidirectional LSTM (BiLSTM) architecture with contextualized ELMo word representations.
- BiLSTM and Multiple Datasets: Athavale et al. (2016) use BiLSTM and multiple datasets to achieve an F1 score of around 77.48% for all NER tags in Hindi.
- Morphological and Phonological Sub-Word Representations: Multilingual approaches explore morphological and phonological sub-word representations to aid NER tasks in languages such as Hindi.
- Typological Features and Machine Learning: C S and Lalitha Devi (2020) propose typological features and machine learning-based approaches for NER in various language families.
- Combined Labelled Data for Indian Language NER: Research demonstrates that training with combined labeled data from multiple languages can improve Indian language NER.

# Creation of Dataset | Overview

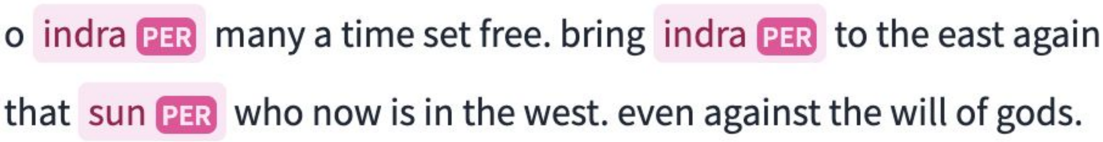
- We used existing parallel corpus to create a Sanskrit NER dataset using Label Transfer Technique
- Parallel Corpuses:
  - Hindi-Sanskrit
  - English-Sanskrit
- Label Transfer
  - Using Hindi and English as the source language run an NER tool for Hindi and English
  - Find the word alignments for the corresponding sentences
  - Transfer the NE labels to Sanskrit using the NER tags created for source language

# Creation of Dataset | Tagset

- We created a dataset with 4 tags:
  - PERSON ('Vālmīki' → 'PERSON' )
  - LOCATION ('Vedas' → 'ORGANIZATION' )
  - ORGANIZATION ( ('Himavat' → 'LOCATION'))
  - O (Other)

# Creation of Dataset | Technique

- For both datasets we have used “Fast Align” tool to calculate alignments in both direction, then took the common alignments.
- For English Dataset
  - After obtaining the alignments we obtained NER tags from ai4bharat/IndicNER - bert-base-multilingual-uncased model



o indra PER many a time set free. bring indra PER to the east again
that sun PER who now is in the west. even against the will of gods.

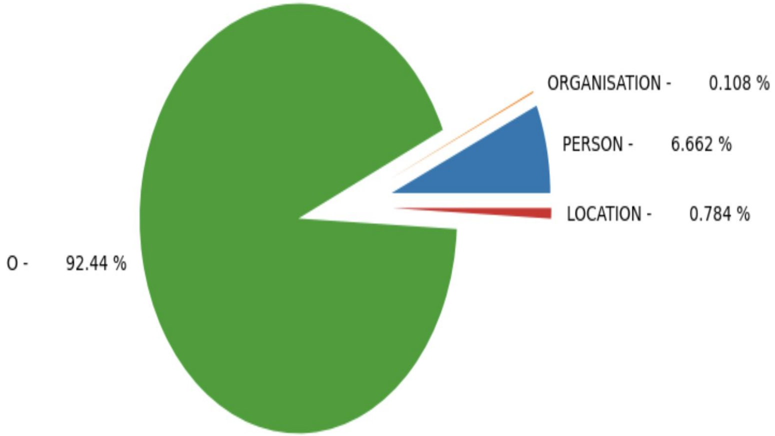
- For Hindi Dataset
  - After obtaining the alignments we obtained NER tags from cfilt/HiNER-collapsed-muril-base-cased -
- As per the align words  we have transferred the tags

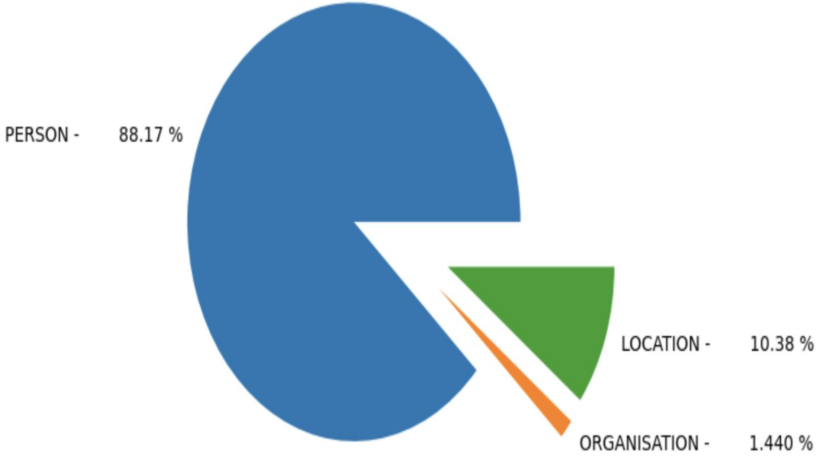
# DataSet Statistics

| Type of Tag | Number of Tokens |
|---|---|
| PERSON | 6509 |
| ORGANISATION | 31 |
| LOCATION | 439 |
| O | 451122 |

- Total Sentences - 45774
- Total Token - 65498
- Total Tags - 4

# DataSet Distribution





# Models

- We trained different models
  - Bi-Directional RNNs
    - Embedding Layer
    - 2 Bi-Directional RNNs
    - Residual Connection
    - Linear Layer
  - Random Forest
  - Adaboost

- GitHub Repository - https://github.com/ujjwalsharmaIITB/Natural-Language-Processing-Semester-1/tree/main/NER%20Project

# Results

| Models | Training Accuracy | Test Accuracy | Validation Accuracy |
|---|---|---|---|
| Bi-Directional RNN (256 units) | 99.4% | 95.59% | 96% |
| Bi-Directional RNN (512 units) | 99.5% | 96.1% | 96% |
| | | | |

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Adaboost | 1.00 | 0.93 | 0.96 |
| Random forest | 1.00 | 0.93 | 0.96 |

# Qualitative Analysis

- Cross-linguistic knowledge transfer from a pre-trained NER model in another language can improve the performance of the NER model in the target language.
- Label transfer can reduce the need for extensive annotation in the target language, saving time and resources.

Some examples of correct alignment

- धृतराष्ट्र - Dhrtarastra
- उवाच - said
- धर्मक्षेत्रे - Kuruksetra,
- कुरुक्षेत्रे - warring
- समवेता - of

# Conclusion

We have tried to adopt a systematic approach to enhance the model's robustness and performance as well as for dataset generation.

We recognize instances where label transfer enhances performance, reinforcing the value of this approach.

We have viewed model development as an iterative process that requires continuous refinement based on feedback and changing data landscapes.

# References

- Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. ***HiNER: A large Hindi Named Entity Recognition Dataset.*** In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4467–4476, Marseille, France. European Language Resources Association.
- ai4bharat/IndicNER , https://huggingface.co/ai4bharat/IndicNER?text=o+indra+many+a+time+set+free.%0D%0Abring+indra+to+the+east+again+that+sun+who+now+is+in+the+west.%0D%0Aeven+against+the+will+of+gods. [30 Nov 2023]
- cfilt/HiNER-collapsed-muril-base-cased , https://huggingface.co/cfilt/HiNER-collapsed-muril-base-cased .[30 Nov 2023]