HiNER: A Large Hindi Named Entity Recognition Dataset

Rudra Murthy[2], Pallab Bhattacharjee[1], Rahul Sharnagat[3], Jyotsana Khatri[1], Diptesh Kanojia[4], Pushpak Bhattacharyya[1]

[1]CFILT Lab, IIT Bombay, India.
[2]IBM IRL, Bangalore, India. [3]Walmart Labs, USA.
[4]Surrey Institute for People-Centred AI, University of Surrey, United Kingdom.
[1]bhattacharjee.pallabg@gmail.com, {jyotsanak, pb}@iitb.ac.in,
[2]rmurthyv@in.ibm.com, [3]rdsharnagat@gmail.com, [4]d.kanojia@surrey.ac.uk

Overview

Motivation

- Fundamental NLP task for a low-resource language scenario.
- Applications to multiple NLP tasks and domains for the Hindi language.
- Gold-standard dataset with human-annotated samples.

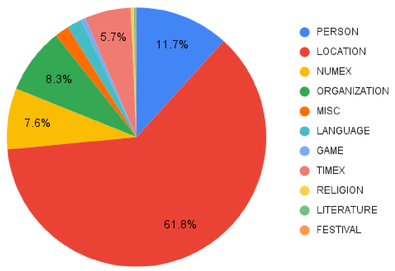
Contributions

- Hindi Named Entity Recognition (NER) dataset **containing > 100k sentences and > 2 million tokens**:
  - **Variant One**: Annotated with (original) 11 NE-Tags.
  - **Variant Two**: Annotated with (collapsed) set of three most-used tags, *viz.*, PER, LOC, ORG.
- Comprehensive performance evaluation across various multi-lingual language models (LMs) with supporting qualitative evaluation.

Dataset statistics

	Ours	Wiki ANN	Fire 2014	IJCNLP 2008
Sentences	**109146**	7000	9622	21833
Tokens	**2220856**	41256	116103	541682
Person	37605	22959	2112	4235
Location	198282	20131	2268	4307
Organization	26509	14204	170	1272

Table 1: HiNER in comparison to existing NER datasets.



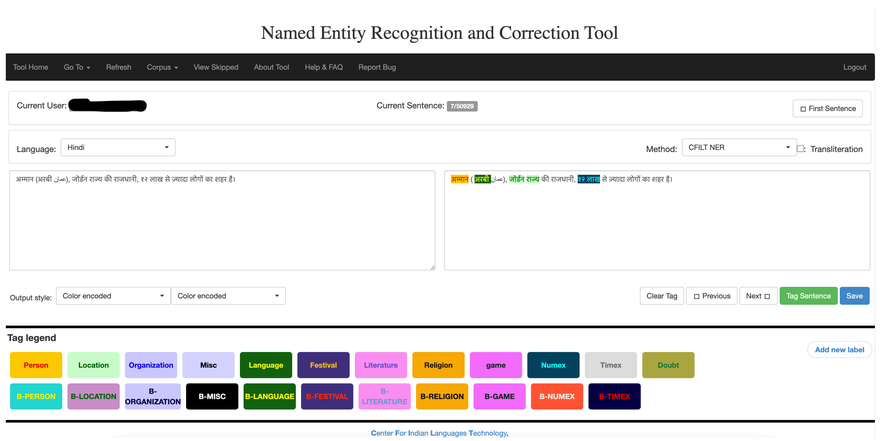
Data Curation

Strategy

- Single Annotator tasked with annotating sentence-level text with 11 NER tags.
- **PHP-MySQL based web-interface** provided to the annotator.
- This interface **utilizes a simple feed-forward neural network** to make token-level NER predictions.
- Annotator is tasked with tagging untagged tokens and **post-editing NE tags** obtained from the neural model.

Tool and Pre-tagged Output

- Backend NER engine trained over FIRE 2013 data which provided quite erroneous suggestions.
- Annotation ambiguity resolved via discussions.



Resource Evaluation

Experiment Setup and Results

- Evaluation performed by **fine-tuning over five pre-trained LMs** which support the Hindi language.
- Hyperparameter tuning over **learning rates**: {1e-3, 3e-3, 5e-3, 1e-4, 3e-4, 5e-4, 1e-5, 3e-5, 5e-5, 1e-6, 3e-6, 5e-6} & **batch sizes**: {8,16,32}.
- We report **mean F1-scores over 5 runs along with their standard deviation** with the best hyperparameters.

	Indic-BERT	mBERT	MuRIL	XLM-R_{base}	XLM-R_{large}
Festival	9.52 ± 11.90	8.57 ± 17.14	0.00 ± 0.00	11.34 ± 14.53	**46.73 ± 23.9**
Game	50.05 ± 8.33	50.92 ± 20.52	40.88 ± 22.96	47.57 ± 10.63	**59.63 ± 7.94**
Language	89.22 ± 1.15	90.07 ± 1.13	90.08 ± 1.02	90.64 ± 0.56	**91.42 ± 0.57**
Literature	21.64 ± 26.12	53.56 ± 10.93	44.23 ± 22.17	40.54 ± 23.39	**56.69 ± 6.32**
Location	94.10 ± 0.56	93.92 ± 0.57	94.81 ± 0.37	94.07 ± 0.76	**94.86 ± 0.40**
Misc	56.14 ± 10.97	61.24 ± 10.99	62.84 ± 4.22	60.38 ± 12.19	**67.86 ± 2.19**
NUMEX	65.56 ± 3.25	67.21 ± 1.50	68.31 ± 1.77	66.72 ± 2.32	**69.10 ± 0.95**
Organization	76.68 ± 1.33	74.81 ± 3.10	78.26 ± 2.46	76.02 ± 2.73	**78.76 ± 1.70**
Person	83.65 ± 0.50	81.10 ± 1.70	84.60 ± 1.30	83.04 ± 0.86	**85.14 ± 0.94**
Religion	65.94 ± 3.20	68.55 ± 7.58	53.43 ± 26.74	67.70 ± 5.78	**72.27 ± 2.68**
TIMEX	80.20 ± 1.11	81.15 ± 1.24	81.17 ± 1.20	79.50 ± 0.85	**80.63 ± 1.05**
Micro	87.44 ± 0.62	87.11 ± 1.01	88.27 ± 0.92	87.36 ± 1.09	**88.73 ± 0.60**
Macro	62.97 ± 5.19	66.46 ± 5.93	63.51 ± 7.36	65.23 ± 6.04	**73.01 ± 3.35**
Weighted	87.25 ± 0.88	87.06 ± 1.28	88.27 ± 1.08	87.29 ± 1.23	**88.78 ± 0.57**

Test Set F1-Score (mean over 5 runs) of pre-trained LMs **on** our **HiNER dataset with original tags**.

	Indic-BERT	mBERT	MuRIL	XLM-R_{base}	XLM-R_{large}
Location	94.33 ± 0.63	94.44 ± 0.24	94.95 ± 0.25	**95.07 ± 0.20**	95.06 ± 0.33
Organization	78.29 ± 1.57	78.42 ± 1.13	79.87 ± 0.81	79.57 ± 0.62	**80.53 ± 0.40**
Person	84.70 ± 0.61	82.20 ± 0.98	**85.66 ± 0.48**	85.18 ± 0.66	85.34 ± 0.54
micro avg	91.37 ± 0.67	91.10 ± 0.34	92.09 ± 0.27	92.06 ± 0.27	**92.20 ± 0.22**
macro avg	85.77 ± 0.91	85.02 ± 0.66	86.83 ± 0.41	86.61 ± 0.40	86.98 ± 0.22
weighted avg	91.34 ± 0.71	91.08 ± 0.35	92.11 ± 0.29	92.11 ± 0.27	92.22 ± 0.22

Test Set F1-Score (mean over 5 runs) of pre-trained LMs **on** our **HiNER dataset with collapsed tags**.

Dataset Links and Documentation

(Documentation/Code) — https://github.com/cfiltnlp/HiNER

Datasets (HiNER-original) — https://huggingface.co/datasets/cfilt/HiNER-original

Datasets (HiNER-collapsed) — https://huggingface.co/datasets/cfilt/HiNER-collapsed