Part-of-speech Tagging for Extremely Low-resource Indian Languages

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya
Department of Computer Science & Engineering, IIT Bombay, Mumbai, India

ACL 2024 Bangkok, Thailand

Introduction

- Modern NLP systems require large amounts of data.
- Extremely low-resource languages (ELRL) have huge data sparsity and are less documented.
- Even for shallow NLP tasks like POS tagging, ELRL do not have datasets.
- Our primary objective is to develop a POS tagger for three extremely low-resourced Indian languages: Angika, Magahi, and Bhojpuri, and compare with Hindi.

Language Profile

Language	Scripts	No. of speakers	Geographical distribution
Angika	Devanagari	15M	Bihar, Jharkhand, West Bengal,Nepal
Bhojpuri	Devanagari	52M	Fiji, Bihar, Uttar Pradesh, Jharkhand
Magahi	Devanagari	14M	Bihar, Jharkhand, Nepal

Languages and their Characteristics

- Our primary languages of interest, Angika, Bhojpuri, Hindi and Magahi, represent the Bihari language group that belongs to the Indo-Aryan language family.
- Script: Devanagari. Word order: Subject-Object-Verb (SOV).
- In Angika & Magahi, object may be absent in simple sentences. E.g.: *Jo kho* (You go and eat)

Experimental Setup

- Zero-shot setting: Train the model using Hindi POS data and evaluate on Angika, Magahi, and Bhojpuri evaluation sets.
- Use F1-score, Precision, and Recall as the performance measures.

Model Description

- We use Transformer-based pre-trained models: MuRIL, RemBERT, and XLM-R-Large. MuRIL variant of BERT, trained on Indic languages only.
- Added a classification layer on top of the encoder output.
- None of these models are trained on Angika, Magahi, and Bhojpuri.

Simple Subword Strategies

- **Look-back (LB)**: Consider the initial token of the word. Update the predictions for all other sub-word tokens with the label predicted for the first sub-word token.

HamRa → *Ham, Ra*
Pron — *Pron, Adp*

HamRa → *Ham, Ra*
Pron — *Pron, Adp/Pron*

- **Look-back-with-score (LBS)**: Consider the token logit scores. Find the sub-word token with the maximum logit score and retrieve its corresponding POS label.

chHaI → *chH, aI*
Aux — *Verb, Aux*
logits — 0.6, 0.7

chHaI → *chH, aI*
Aux — *Verb, Aux*

Results using Subword Strategies

		Hindi			Angika			Magahi			Bhojpuri		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Hindi FT	MuRIL	0.93	0.94	0.93	0.66	0.72	0.69	0.74	0.78	0.76	0.78	0.80	0.79
	XLM-R large	0.93	0.93	0.93	0.68	0.69	0.69	0.73	0.74	0.74	0.76	0.77	0.76
	RemBERT	0.93	0.94	0.94	0.67	0.71	0.69	0.75	0.77	0.76	0.76	0.77	0.76
	AVG	0.93	0.94	0.93	0.67	0.71	0.69	0.74	0.76	0.75	0.76	0.77	0.77
Look-back	MuRIL	**0.95**	0.94	0.94	**0.80**	**0.77**	**0.77**	**0.84**	**0.82**	**0.83**	**0.85**	**0.83**	**0.84**
	XLM-R large	**0.95**	**0.95**	**0.95**	0.78	0.73	0.74	0.81	0.76	0.77	0.80	0.75	0.75
	RemBERT	0.94	0.94	0.94	0.79	0.73	0.74	0.82	0.79	0.80	0.82	0.78	0.79
	AVG	0.95	0.94	0.94	0.79	0.74	0.75	0.82	0.79	0.80	0.83	0.79	0.79
Look-back-with-score	MuRIL	**0.95**	**0.95**	**0.95**	**0.80**	0.76	**0.77**	0.83	0.81	0.82	**0.85**	0.82	0.83
	XLM-R large	**0.95**	0.94	0.94	0.79	0.74	0.75	0.83	0.80	0.80	0.83	0.80	0.80
	RemBERT	**0.95**	**0.95**	**0.95**	**0.80**	0.75	0.76	0.83	0.81	0.81	0.84	0.81	0.81
	AVG	0.95	0.95	0.95	0.80	0.75	0.76	0.83	0.81	0.81	0.84	0.81	0.81

- MuRIL outperforms all other mPLMs across all languages, demonstrating the effectiveness of family-specific training.
- On average, Look-back-with-score outperform Look-back ($+1\%$ increase in F1-score across all languages and models).

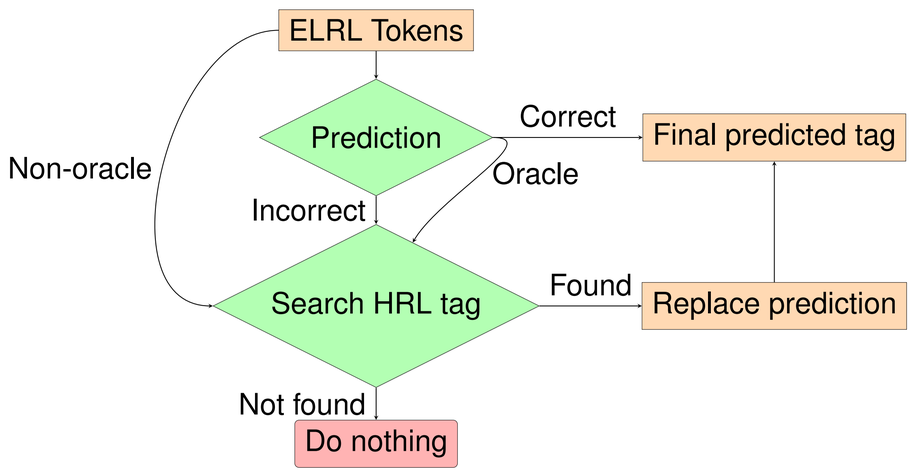
Tag-level Performance

	Hindi FT					Look-back-with-score				
Tags	Hindi	Angika	Magahi	Bhojpuri	AVG	Hindi	Angika	Magahi	Bhojpuri	AVG
ADJ	0.89	0.68	0.74	0.76	0.73	0.95	0.73	0.74	0.76	**0.74**
ADP	0.96	0.82	0.94	0.94	**0.90**	0.98	0.82	0.94	0.94	**0.90**
ADV	0.95	0.36	0.51	0.48	**0.45**	0.95	0.36	0.49	0.48	0.44
AUX	0.96	0.78	0.84	0.81	**0.81**	0.98	0.78	0.84	0.82	**0.81**
CCONJ	0.93	0.70	0.87	0.88	0.82	0.97	0.89	0.87	0.88	**0.88**
DET	0.8	0.44	0.47	0.59	0.50	0.81	0.43	0.58	0.55	**0.52**
NOUN	0.93	0.83	0.85	0.87	**0.85**	0.95	0.83	0.85	0.87	**0.85**
NUM	0.9	0.75	0.69	0.74	0.73	0.92	0.75	0.69	0.82	**0.75**
PART	0.91	0.47	0.79	0.75	**0.67**	0.88	0.44	0.79	0.74	0.66
PRON	0.93	0.69	0.7	0.63	0.67	0.96	0.78	0.7	0.75	**0.74**
PROPN	0.85	0.5	0.57	0.49	0.52	0.87	0.64	0.57	0.59	**0.60**
PUNCT	0.98	0.99	0.98	1.00	**0.99**	0.98	0.99	0.98	1.00	**0.99**
SCONJ	0.83	0.63	0.68	0.63	**0.65**	0.83	0.63	0.68	0.63	**0.65**
SYM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
VERB	0.95	0.71	0.75	0.82	**0.76**	0.97	0.74	0.75	0.82	**0.76**

- After applying look-back-with-score on average, Conjunction, Pronoun, and Proper-noun tags get the highest performance gains.
- Angika: conjunctions, proper-nouns, and pronouns get $19\%, 14\%,$ and 9% gains in F1-scores, respectively.
- Bhojpuri: pronouns, proper-nouns, and numbers get $12\%, 10\%,$ and 8% gains in F-1 scores, respectively.

Using High-resource Parallel Data

- **Oracle**: Compare our model's POS predictions for ELRL with ground-truth tags and correct mismatched tokens' tags with HRL token's true tags.
- **Non-oracle**: Directly replacing ELRL tokens' tags with HRL token's true tags.



Results

		Hindi			Angika			Magahi			Bhojpuri		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Hindi FT	MuRIL	**0.93**	**0.94**	0.93	0.66	0.72	0.69	0.74	0.78	0.76	0.78	0.80	0.79
	XLM-R large	**0.93**	0.93	0.93	0.68	0.69	0.69	0.73	0.74	0.74	0.76	0.77	0.76
	RemBERT	**0.93**	**0.94**	**0.94**	0.67	0.71	0.69	0.75	0.77	0.76	0.76	0.77	0.76
	AVG	**0.93**	**0.94**	**0.94**	0.67	0.71	0.69	0.74	0.76	0.75	0.76	0.77	0.77
Oracle	MuRIL	-	-	-	**0.81**	**0.78**	**0.79**	**0.85**	**0.84**	**0.84**	**0.87**	**0.85**	**0.85**
	XLM-R large	-	-	-	0.80	0.76	0.76	0.82	0.79	0.79	0.83	0.80	0.80
	RemBERT	-	-	-	0.80	0.75	0.76	**0.85**	0.83	0.83	0.85	0.84	0.84
	AVG	-	-	-	0.80	0.76	0.77	0.84	0.82	0.82	0.85	0.83	0.83
Non-oracle	MuRIL	-	-	-	0.70	0.64	0.64	0.73	0.68	0.69	0.75	0.71	0.72
	XLM-R large	-	-	-	0.57	0.51	0.49	0.64	0.60	0.57	0.67	0.63	0.60
	RemBERT	-	-	-	0.58	0.50	0.53	0.65	0.59	0.62	0.66	0.61	0.63
	AVG	-	-	-	0.62	0.55	0.55	0.67	0.63	0.63	0.69	0.65	0.65

- Replacing wrongly predicted token tags with high resource language tags improves the overall tagging performance.
- Directly replacing token tags with high resource language tags degrades performance to lower than Hindi-FT.

Analysis

- MURIL consistently outperforms other models across all test languages.
- On average, look-back-with-score outperforms look-back w.r.t. F1-scores across all languages.
- Using a parallel corpus in a related HRL like Hindi improves the overall tagging performance, primarily when focusing on wrongly predicted tags.
- Without prior knowledge of a wrongly predicted tag, replacing it with HRL tags hurts performance.
- With look-back and look-back-with-score, we find notable performance gains for tags like pronouns, numbers, conjunctions, adjectives, and proper nouns.

Conclusions

- We offer one of the first NLP resources for Angika and Magahi following UD guidelines.
- We provide state-of-the-art POS tag baselines by fine-tuning multilingual PLMs with Hindi data.
- We find that pretrained tokenizers adversely affect cross-lingual transfer. We address this for POS-tagging with a simple look-back scheme.

More Details

Paper, code, and data available at `https://github.com/snjev310/acl-24-pos`

Paper

Code & Dataset

Acknowledgements

The first author would like to gratefully acknowledge a Ph.D. grant from the TCS Research Foundation.

References

- C Chandramouli and Registrar General. 2011. Census of india 2011. Provisional Population Totals. New Delhi: Government of India, pages 409–413
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multi- lingual representations for indian languages. arXiv preprint arXiv:2103.10730.