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

Sanjeev Kumar, Preethi Jyothi, Pushpak Bhattacharyya
Computer Science and Engineering, IIT Bombay, India
{sanjeev, pjyothi, pb}@cse.iitb.ac.in

Abstract

Modern natural language processing (NLP) systems thrive when given access to large datasets. However, a large fraction of the world's languages are not privy to such benefits due to sparse documentation and inadequate digital representation. This is especially true for Indian regional languages. As a first step towards expanding the reach of NLP technologies to extremely low-resource Indian languages, we present a new parallel part-of-speech (POS) evaluation dataset for Angika, Magahi, Bhojpuri and Hindi. Angika, Magahi, Bhojpuri, along with the more well-known Hindi, are all languages spoken in the Indian states of Bihar, Jharkhand and West Bengal. Ours is notably the first NLP resource, even for a shallow NLP task like POS-tagging, for Angika. We establish POS-tagging baselines using state-of-the-art multilingual pretrained language models (PLMs) finetuned on Hindi data, and show zero-shot evaluations on the other three languages. While all four languages use the same Devanagari script, pretrained tokenizers underperform in zero-shot on the three languages. We propose a simple look-back fix to address the tokenization challenge yielding F1-score improvements of up to 8% on Angika, and show how it comes very close to an oracle setting when the underlying Hindi word is known (and can be accurately tokenized).

1 Introduction

India is a multilingual country with more than 1369 languages and five main language families (Office of the Registrar General Census Commissioner, 2022). While the Indian constitution officially recognizes 22 languages, numerous others face a battle for survival. English is spoken by only roughly 10% of the population (Office of the Registrar General Census Commissioner, 2022) in India; the majority prefers their diverse regional languages deeply rooted in cultural heritage. Building technologies for regional Indian languages is important to ensure inclusivity across user groups and empower people for everyday interactions.

While there has been progress towards promoting multilinguality and linguistic diversity across Indian languages with tools like IndicNLPSuite (Kakwani et al., 2020) and multilingual corpora such as Common Crawl Oscar Corpus (Wenzek et al., 2019), PMIndia (Haddow and Kirefu, 2020), and Samanantar (Ramesh et al., 2021), there is almost no representation of low-resource Indian languages like Angika in these resources.

In this work, we focus on three very low-resource Indian languages Angika, Magahi and Bhojpuri and create parallel POS-tagging evaluation corpora for these three languages, consistent with Universal Dependencies (UD) guidelines. These are the first UD-compliant datasets for Angika and Magahi. We also create a Hindi POS-tagging dataset that is parallel to the data in the three languages. Hindi is the closest high-resource Indian language that is related to Bhojpuri, Angika and Magahi. This allows us to carefully examine the cross-lingual performance gap compared to Hindi, whether transfer from a related high-resource language like Hindi is possible, and challenges related to tokenization that affect zero-shot performance on the low-resource languages. We propose a simple look-back scheme that circumvents most errors that stem due to suboptimal tokenization for the three low-resource languages. To encourage further work on these languages, we publicly release our new dataset and code to reproduce all our experiments [1].

2 Related Work

Shallow NLP tasks such as POS tagging for low-resource languages have been studied fairly extensively in prior work, many of which have explored cross-lingual transfer learning techniques. Fang

[1] https://www.github.com/snjev310/acl-24-pos

Findings of the Association for Computational Linguistics ACL 2024, pages 14422–14431

and Cohn (2016) achieve successful POS tagging, especially in low-resource languages (like Malagasy and Kinyarwanda), by combining word alignment with gold-standard data. Kim et al. (2017) implemented a BiLSTM model that utilizes word and character embeddings to transfer knowledge without relying on parallel corpora. Another approach by Huck et al. (2019) introduces zero-shot tagging, by projecting annotations from a related high-resource language, such as Russian for Ukrainian. More recently, Dione et al. (2023) demonstrated significant improvements in POS tagging by using multilingual pretrained LMs trained on typologically diverse African languages. Recent work has also employed modular learning (Lin et al., 2019; Artetxe et al., 2020; Pfeiffer et al., 2020; Min et al., 2023) techniques with multilingual pretrained language models (PLMs) to enable effective cross-lingual transfer to low-resource languages. Prior work on POS tagging for various low-resource Indo-Aryan languages has mainly utilized classical NLP techniques and do not leverage PLMs. For example, Saharia et al. (2009) developed a POS tagger for Assamese and Basit and Kumar (2019) for Awadhi, and both works used an HMM model.

Among the four languages of interest, other than Hindi, there is an existing UD-compliant POS-tagging evaluation dataset for Bhojpuri (Ojha and Zeman, 2020). However, it only has 268 sentences. Overall, there is a strong need for NLP datasets covering extremely low-resource Indian languages (such as Angika). We think it is also useful to develop parallel datasets along with relatively higher-resource languages (such as Hindi) that share common geographical boundaries, word order, script, and language family. This helps understand how cross-lingual transfer can be more effectively utilized from high-resource languages.

3 Data Collection and Annotation

We created a parallel corpus comprising 708 evaluation sentences each for Hindi, Angika, Bhojpuri, and Magahi. We used some Hindi monolingual data from Kunchukuttan et al. (2018) and translated into Angika, Magahi, and Bhojpuri. Some Angika monolingual data was extracted from the webonary[2] Angika dictionary and translated into Hindi and the other two languages. The UD dataset primarily draws data from the news domain, resulting in a higher frequency of named entities belonging to the noun class. To improve diversity, we incorporated sentences of daily conversations and regional stories. Additionally, we included a few frequently used sentences by native speakers. Issues in the webonary Angika dictionary such as word misspellings and missing sentences were manually addressed. Languages were translated by respective language annotators and verified by native speakers. Table 4 in Appendix A.2 provides more details for all four languages and Appendix A.4 provides more details about our measures for quality control in the annotations.

[2] https://www.webonary.org/angika/

4 Experiments and Results

We provide POS tagging baselines using multilingual PLMs in a zero-shot setting. We fine-tune an Indic-specific multilingual PLM (MURIL (Khanuja et al., 2021)) on Hindi UD POS-tagging data consisting of 13K sentences. We refer to this model as “Hindi FT". We also fine-tuned on other massively multilingual PLMs, XLM-R-large (Conneau et al., 2020) and RemBERT (Chung et al., 2020)), as a comparison to MURIL. All three PLMs have been pretrained on around 100-110 languages, with the exception of MuRIL, which is exclusively pretrained on 12 major Indic languages. However, none of these models are pretrained on any of the low-resource languages that we explore in this work. More details regarding the experimental setup can be found in Appendix A.5.

4.1 Baseline Models

Table 1 presents the zero-shot POS tagging results for all three low-resource languages (under “Hindi FT"). MuRIL has a slight advantage over XLM-R and RemBERT in the zero-shot setting. MuRIL is significantly smaller in size, compared to XLM-R-large and RemBERT. However, it performs best on the low-resource languages possibly by benefiting from its Indic-only pretraining, as opposed to the other PLMs that will have language interference from many non-Indic languages.

We also evaluate MuRIL on publicly available UD Hindi test data to validate our model setup. On the UD Hindi test data, MuRIL achieves an F1-score of 0.96, which is comparable to the F1-score we obtained on our Hindi evaluation set (0.93). This validates our model setup and acts as a sanity check for the quality of our Hindi data.

		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

Table 1: POS tagging Precision (**P**), Recall (**R**), and F1-score (**F1**) of three PLMs evaluated on our parallel dataset using zero-shot, look-back, and look-back-with-score methods. Results range from 0 to 1 and are averaged across five random seeds.

4.2 Tokenization Inconsistencies

In the case of extremely low-resource languages, a pretrained tokenizer tends to break words into multiple sub-word tokens. In our experiments, we predict POS tags for each token. Thus, sub-optimal tokenization can result in poor quality predictions. For extremely low-resource languages, words may be split all the way down into individual characters which makes the POS predictions even noisier. In Section 4.3, we propose a simple look-back scheme to alleviate some of the issues that stem from poor tokenization. In Section 4.4, we analyze how much we could make up for tokenization challenges in an oracle setting if we had access to parallel data in Hindi.

4.3 Investigating the impact of sub-word

When a tokenizer breaks down words into sub-words, these sub-words can get different tags. Poor tokenization that leads to over-fragmentation exacerbates this problem. In our quantitative dataset analysis, we observed that in Angika, Magahi and Bhojpuri around 45% of words were split into 2, 3, and 4 sub-word tokens. To address this challenge, we introduce two simple techniques: look-back and look-back-with-score.

Look-back. We substitute the POS tags corresponding to all the tokens in a word with the POS tag of the first token. This approach was chosen because the first split token closely relates to the word in a higher-resource language, preserving meaning and POS tags. For example, an Angika word “dEkhAibae" (will see), splits into “dEkh” and “ibae” here, the word “dEkh” (see) is related to Hindi. The results of this look-back approach are shown in Table 1. We find significant improvements in performance across all three low-resource languages, most notably for Angika.

Look-back-with-score. Each token produces logit values corresponding to the tag distributions. If a word W is split into n sub-words $(w_1, w_2, \ldots, w_n)$ with corresponding logits $(l_1, l_2, \ldots, l_n)$, we find the sub-word token with the maximum logit score and retrieve its corresponding POS label. We replace the POS tags of all other sub-word tokens in that word with the POS tag of the maximum-scoring token.

When comparing look-back and look-back-with-score methods in Table 1, the results suggest that, on average, look-back-with-score outperforms look-back (+1% increase in F1-score across all languages and all models). Our initial observations indicate that tokenizers generally split words into tokens that are seen in the vocabulary of high-resource languages, and thus replacing the tag of subsequent tokens with the tag of the first token is a reasonable strategy to mitigate token-level inconsistencies. However, the “look-back with score” approach utilizes information from tokens within a word, and this turns out to offer a small but consistent advantage across all three languages. The look-back-with-score approach is less effective for MuRIL compared to look-back. This may be because MuRIL, primarily trained on Indian languages, aligns initial split tokens more closely with Hindi, resulting in a higher confidence level for the first token than for other split tokens.

4.4 Using Hindi Parallel Data

Here, we assess the effectiveness of having access to a parallel corpus of a relatively higher resource

		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

Table 2: POS tagging Precision (**P**), Recall (**R**), and F1-scores (**F1**) of three PLMs evaluated on our parallel data using zero-shot ('Hindi FT'), oracle and non-oracle methods. Results range from 0 to 1 and are averaged across five random seeds.

language, which shares the script, word order, and geographical boundaries and can assist in POS tag predictions for extremely low-resource languages. For this analysis, we have introduced two methods: the oracle setting, where we have oracle knowledge of incorrectly predicted POS tags, and the non-oracle setting, where we lack any prior knowledge of incorrectly predicted POS tags.

Oracle setting. We compare our model's predictions for Angika, Magahi and Bhojpuri with the ground-truth POS tag sequences. For mismatches, we utilize the parallel corpus in Hindi to identify the corresponding Hindi token and the most frequently occurring tag for the aligned Hindi token. This tag replaces our model predictions in the other three low-resource languages.

Non-oracle setting. We substitute the predicted tags of a token with the original tags from the parallel corpus. This involves choosing the tag with the highest frequency for an aligned Hindi token in the parallel corpus. In this setting, we achieve F1-scores of 0.67, 0.63, and 0.65 for Angika, Magahi, and Bhojpuri, respectively, which is significantly lower than the scores obtained in the zero-shot and oracle settings.

Table 2 presents the results of both approaches. When comparing the oracle and non-oracle results with the zero-shot approach, we observe that having access to a parallel corpus in a comparatively higher-resource language improves the overall tagging performance, primarily when focusing on wrongly predicted tags. However, since we do not have apriori information of whether our predicted tag is correct or not, the non-oracle approach of always replacing it with the Hindi token's most frequent POS tag results in a notable decline in performance.

4.5 Tag-level Analysis

Table 1 shows significant improvements in performance using look-back-with-score over the zero-shot baseline. These methods provide notable performance gains for tags like pronouns, proper nouns, conjunctions, adjectives, and numbers. We illustrate how a simple strategy like look-back significantly helps with an example. Consider the Angika sentence "HamMe aArU tOI miLika duGo aAma kHaIlIye" (Translation: You and I ate two mangoes together). While a pretrained tokenizer correctly identifies "hamMe" as two pronouns in Angika, it might misrecognize their individual tags ("ham" as pronoun, "Me" as adposition). The look-back method leverages contextual information to rectify these errors by correctly predicting both tokens as pronouns. This results in an improvement over the zero-shot setting. Similar morphological structures are observed with numerals like "duGo" (meaning "two"). All these languages employ a classifier as a bound morpheme, with "Go" as the specific marker for numbers. A broader analysis of the linguistic properties causing errors in the zero-shot setting is discussed in 4.6.

4.6 Error Analysis

Table 3 shows tag-wise F1-scores for different languages using MuRIL in zero-shot and look-back-with-score settings. Subordinate conjunctions, pronouns, proper nouns, particles, determiners, and adverbs exhibit lower F1-scores compared to other tags. Adjectives and numbers also show lower scores than Hindi and the total occurrences of tags

	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**

Table 3: Tag-wise F1-scores using MuRIL for POS tags across all languages for the zero-shot baseline ('Hindi FT') and 'look-back-with-score' method. Average (AVG) is calculated using Angika, Magahi, and Bhojpuri scores.

in the dataset, particularly when compared to particles and determiners. These disparities can be attributed to complexities in word formations and a few other linguistic nuances detailed below.

Tag confusion between subordinate conjunction, adposition, and pronoun. In Angika, Magahi and Bhojpuri, pronouns, adpositions and subordinate conjunctions are written similarly. For e.g.:
Hindi: kyA tum kal aAogI?
Angika: kI toI kaAl aAiybae?
Bhojpuri: kA tu kal aAiyebu?
Magahi: kA tu kal aAibhe?
English: Will you come tomorrow?
In this example, kyA, kI, and kA are pronouns, but kI and kA are written as Hindi adposition and subordinate conjunctions.

Classifiers. Angika, Bhojpuri, and Magahi languages often make use of classifier markers (Tho, Go) to accompany numbers (e.g., Ek (one) and dU (two)). Classifier markers in these languages are: To, Te, go, Ke, etc. Hindi does not exhibit this property. For e.g.:
Hindi: Ek sajJan.
Angika: EkTa SajJanA.
Bhojpuri: EGo sajJanA.
Magahi: EaGo sajJanA.
English Translation: One gentleman.

Non-ergative construction. Hindi is an ergative language, while Angika, Bhojpuri, and Magahi are non-ergative. In Hindi, when the subject is designated with [ne] (oblique case), the transitive verb agrees with the object in terms of person, number, and gender. For e.g.:
Hindi: Mohan ne kitaB padi.
Angika: Mohane kitaB padhalkae.
Bhojpuri: Mohane kitaB padhlAs.
Maghai: Mohane kitaB padalAi.
English Translation: Mohan read the book.

5 Conclusion

This paper introduces a UD-compliant parallel POS tag dataset for three extremely low-resource Indian languages: Angika, Magahi and Bhojpuri. This work contributes one of the first NLP resources for these languages and is a first step towards addressing their under-representation in the digital landscape. We provide state-of-the-art POS baselines by fine-tuning multilingual PLMs with Hindi data. We find that pretrained Indic tokenizers adversely affect cross-lingual transfer to Angika, Magahi and Bhojpuri which we largely address with a simple look-back scheme.

Acknowledgements

The first author would like to gratefully acknowledge a Ph.D. grant from the TCS Research Foundation to support his research on extremely low-resource Indian languages. We would like to thank the anonymous reviewers for their helpful comments.

Limitations

1. While we cover three low-resource languages, many others like Bajjika, Surajpuri, and Maithili, which share geographical boundaries with our languages of interest, have not been included.

2. Due to the smaller size of our dataset compared to larger POS datasets, we do not have a lot of diversity across domains. We note that the POS dataset and the observations in this work may not be applicable to all domains, such as speech transcripts or conversational data.

Ethics Statement

We would like to emphasize our commitment to upholding ethical practices throughout this work. We aimed to ensure that human annotators received a fair compensation for their annotation efforts and was commensurate with the time and effort invested in their work.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Abdul Basit and Ritesh Kumar. 2019. Towards a part-of-speech tagger for awadhi: Corpus and experiments. In *(IJACSA) International Journal of Advanced Computer Science and Applications*.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

Ronald Cardenas, Ying Lin, Heng Ji, and Jonathan May. 2019. A grounded unsupervised universal part-of-speech tagger for low-resource languages. 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)*, pages 2428–2439, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Richard Oliver Collin. 2010. Ethnologue. *Ethnopolitics*, 9(3-4):425–432.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

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: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2023. Analyzing dataset annotation quality management in the wild. *arXiv preprint arXiv:2307.08153*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

India (ORGI) Office of the Registrar General Census Commissioner. 2022. Census of India 2011 - LANGUAGE ATLAS - INDIA. https://censusindia.gov.in/nada/index.php/catalog/42561. [Accessed 05-06-2024].

Atul Kr. Ojha and Daniel Zeman. 2020. Universal dependency treebanks for low-resource indian languages: The case of bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Navanath Saharia, Dhrubajyoti Das, Utpal Sharma, and Jugal Kalita. 2009. Part of speech tagger for assamese text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 33–36.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

A Appendix

A.1 Extremely low resource languages

An extremely low-resource language has relatively few or no resources available. Compared to other languages, Indian regional languages have limited resources. These extremely low-resource languages are often considered less popular, poorly documented, under-resourced, minority, or underdigitized due to their scarce resources. Much of the data and documentation for these languages remain unpublished, exist only in print, or are extremely limited. Consequently, accessing and utilizing raw text in an extremely low-resource language is challenging.

A.2 Language description

Table 4 provides information on the languages of interest, including details such as scripts, number of speakers, data size, and geographical distributions.

A.3 Languages and their characteristics

Our primary languages of interest are Angika, Bhojpuri, Hindi, and Magahi, spoken in the Indian states of Bihar, Jharkhand, West Bengal, and some parts of Nepal. Table 4 provides an overview of the selected languages. These selected languages, representing the Bihari language group, belong to the Indo-Aryan language family (Collin, 2010). They all use the Devanagari script, having transitioned from their individual writing scripts. All of these languages demonstrate tonal characteristics. As far as morphosyntax is concerned, a separate morpheme marks the plural instead of a suffix. All the mentioned languages feature a bound morpheme acting as a classifier. In addition to singular and plural forms, Angika, Magahi, and Bhojpuri contain equal, honorific, and non-honorific variations of second-person personal pronouns. Derived adjectives are present in all languages. While Angika and Bhojpuri offer all three tenses, Bhojpuri lacks morphological availability for the present tense. Angika exhibits simple or

Language	Scripts	No. of speakers	No. of sentences	Geographical distribution
Angika	Devanagari	15M	708	Bihar, Jharkhand, West Bengal,Nepal
Bhojpuri	Devanagari	52M	708	Fiji, Bihar, Uttar Pradesh, Jharkhand, Chhattisgarh, Madhya Pradesh
Magahi	Devanagari	14M	708	Bihar, Jharkhand, Nepal

Table 4: Data statistics showing language, writing script, number of native speakers, geographical distribution, and number of sentences in our POS-tagging corpus across all four languages.

continuous tenses, and Magahi distinguishes between future and non-future tenses. The word order for all languages is Subject-Object-Verb (SOV).

Closeness to Hindi Hindi, Angika, Bhojpuri, and Magahi share similar word order (SOV) and scripts and belong to the Indo-Aryan family. They differ significantly in number, gender, tense, aspect, mood, and case markers. For example, Hindi verbs undergo extensive conjugation for person, number, tense, and mood. In contrast, Angika and Magahi verbs follow distinct conjugational patterns, leading to variations in agreement and verb forms within sentences. Furthermore, the structure and placement of relative clauses differ between the two languages. Despite all languages utilizing the Devanagari script, identical words may convey different meanings. For example, consider the Hindi sentence “hamen bachaav ke baare mein kuchh bhee paravaah nahin thee” (We did not care about saving anything), which translates to Angika as “hammae sinee ka kuchchhoo bachaay ro baare mein paravaay nai chelai.” Here, the pronoun "hamen" (we) in Hindi is expressed in Angika as "hammae sinee ka" (we). While in Hindi, it represents the first-person singular form, in Angika, it denotes the first-person plural form.

A.4 Quality control

To assess the quality of the POS annotation task, we abstained from computing automatic inter-annotator agreement metrics like Fleiss Kappa (Fleiss, 1971). For sequence labelling datasets, dataset creators did not compute agreement as it relied on token level span (Klie et al., 2023). Brandsen et al. (2020) claims that per-token agreement in sequence labelling presents challenges, as annotators label sequences instead of individual tokens, diluting the measure’s ability to capture the essence of the task. Additionally, nouns dominate the labelled data, creating an imbalanced dataset that could skew results. Considering these challenges, we opted for in-person discussions with annotators to reach a consensus on correct annotations for each word in the sentence. The manual, in-person approaches to quality control are suitable and provide reassurance regarding the high quality of the annotation (Cardenas et al., 2019). In cases of disagreement, annotators collaborated with language experts to resolve the issue. For Hindi, sentence-level annotation achieved over 90% agreement; for Angika, Bhojpuri, and Magahi, sentence-level annotation reached over 88% agreement. After quality control, our corpus is the most extensive parallel corpus for Hindi, Angika, Bhojpuri, and Magahi within the UD dataset, where test sets typically involve 300 sentences.

A.5 Experiment setup

For fine-tuning the PLMs, we employed a batch size of 16, a learning rate of 2e-5, and a weight decay of 0.01, and we conducted the experiments over 10 epochs. The computations were performed using Nvidia A100 GPU.

Tags	**Hindi**	**Angika**	**Magahi**	**Bhojpuri**	**AVG**
ADJ	0.89	0.68	0.74	0.76	0.73
ADP	0.96	0.82	0.94	0.94	0.90
ADV	0.95	0.36	0.51	0.48	0.45
AUX	0.96	0.78	0.84	0.81	0.81
CCONJ	0.93	0.70	0.87	0.88	0.82
DET	0.78	0.43	0.48	0.55	0.49
INTJ					
NOUN	0.93	0.83	0.85	0.87	0.85
NUM	0.9	0.75	0.69	0.74	0.73
PART	0.91	0.47	0.79	0.75	0.67
PRON	0.93	0.69	0.7	0.63	0.67
PROPN	0.85	0.5	0.57	0.49	0.52
PUNCT	0.98	0.99	0.98	1.00	0.99
SCONJ	0.83	0.63	0.68	0.63	0.65
SYM	1.00	1.00	1.00	1.00	1.00
VERB	0.95	0.71	0.75	0.82	0.76
X					
-					

Table 5: F1-scores for POS tags across all languages. The average is calculated over Angika, Magahi, and Bhojpuri only. All scores are for MuRIL in the zero-shot setting, as it outperforms RemBERT and XLM-R-Large in this scenario.

Tags	Hindi	Angika	Magahi	Bhojpuri	AVG
ADJ	0.92	0.73	0.74	0.78	0.75
ADP	0.98	0.82	0.94	0.94	0.90
ADV	0.96	0.40	0.45	0.48	0.44
AUX	0.98	0.80	0.86	0.84	0.83
CCONJ	0.96	0.92	0.88	0.89	0.90
DET	0.8	0.44	0.47	0.59	0.50
INTJ					
NOUN	0.96	0.87	0.86	0.89	0.87
NUM	0.91	0.76	0.68	0.69	0.71
PART	0.90	0.47	0.76	0.72	0.65
PRON	0.96	0.76	0.8	0.76	0.77
PROPN	0.88	0.64	0.65	0.61	0.63
PUNCT	0.99	0.99	1.00	1.00	1.00
SCONJ	0.85	0.53	0.71	0.70	0.65
SYM	1.00	1.00	1.00	1.00	1.00
VERB	0.97	0.78	0.81	0.86	0.80
X					
-					

Table 6: F1-scores for POS tags across all languages. The average is calculated over Angika, Magahi, and Bhojpuri only. All scores are for MuRIL in the look-back setting.

Tags	Hindi	Angika	Magahi	Bhojpuri	AVG
ADJ	0.95	0.73	0.74	0.76	0.74
ADP	0.98	0.82	0.94	0.94	0.90
ADV	0.95	0.36	0.49	0.48	0.44
AUX	0.98	0.78	0.84	0.82	0.81
CCONJ	0.97	0.89	0.87	0.88	0.88
DET	0.81	0.43	0.58	0.55	0.52
INTJ					
NOUN	0.95	0.83	0.85	0.87	0.85
NUM	0.92	0.75	0.69	0.82	0.75
PART	0.88	0.44	0.79	0.74	0.66
PRON	0.96	0.78	0.7	0.75	0.74
PROPN	0.87	0.64	0.57	0.59	0.60
PUNCT	0.98	0.99	0.98	1.00	0.99
SCONJ	0.83	0.63	0.68	0.63	0.65
SYM	1.00	1.00	1.00	1.00	1.00
VERB	0.97	0.74	0.75	0.82	0.76
X					
-					

Table 7: F1-scores for POS tags across all languages. The average is calculated over Angika, Magahi, and Bhojpuri only. All scores are for MuRIL in the look-back-with-score setting.