Evaluating Cross Lingual Transfer for Morphological Analysis: a Case Study of Indian Languages

Siddhesh Pawar*
Google Research
siddheshmp@google.com

Partha Talukdar
Google Research
partha@google.com

Pushpak Bhattacharyya
IIT Bombay
pb@cse.iitb.ac.in

Abstract

Recent advances in pretrained multilingual models such as Multilingual T5 (mT5) have facilitated cross-lingual transfer by learning shared representations across languages. Leveraging pre-trained multilingual models for scaling morphology analyzers to low-resource languages is a unique opportunity that has been under-explored so far. We investigate this line of research in the context of Indian languages, focusing on two important morphological subtasks: root word extraction and tagging morphosyntactic descriptions (MSD), viz., gender, number, and person (GNP). We experiment with six Indian languages from two language families (Dravidian and Indo-Aryan) to train a multilingual morphology analyzers for the first time for Indian languages. We demonstrate the usability of multilingual models for few-shot cross-lingual transfer through an average 7% increase in GNP tagging in a cross-lingual setting as compared to a monolingual setting through controlled experiments. We provide an overview of the state of the datasets available related to our tasks and point-out a few modeling limitations due to datasets. Lastly, we analyze the cross-lingual transfer of morphological tags for verbs and nouns, which provides a proxy for the quality of representations of word markings learned by the model.

1 Introduction

Morphology analysis is the first step of processing in the classical NLP pipeline. Even in the transformer era, wherein the entire NLP pipeline is replaced with a transformer, the use of morphological segmentation for tokenization instead of statistical subword tokenization has been shown to produce better embeddings, especially for morphologically rich languages (Nzeyimana and Rubungo, 2022). The statistical subword tokenization used in tokenizers such as wordpiece cannot capture morphological alternations (e.g. wordpiece doesn't treat contextual allomorphs as related) and non-concatenative morphology (Klein and Tsarfaty, 2020).

One of the tasks that we analyze in our work is root word extraction, which forms an integral component of morphologically informed segmentation. A morphology analyzer can also help speed up language documentation efforts for endangered languages, Moeller et al. (2020) leveraged interlinear glossed text to generate unseen forms of inflectional paradigm using a morphology analyzer. Availability of morphological information can also benefit various downstream tasks such as parsing (Seeker and Çetinoğlu, 2015), machine translation (Tamchyna et al., 2017), language modeling (Park et al., 2021), etc. Our scope of this work is inflectional and concatenative morphology. We also envision our work to be used in bias-aware machine translation, especially from morphologically poor languages to morphologically richer languages. For example, if we want to translate the sentence "My friend was a doctor" to Hindi, we would ideally prefer to have both masculine and feminine translations "Mera dost doctor tha" (masculine) and "Meri dost doctor thi"(feminine), as English sentence has no mention of gender and for Hindi, the gender markers are present on verbs (tha\thi) and pronouns (mera\meri).

Although high-quality morphology analyzers have been built for some Indian languages, they are either rule-based such as Agarwal et al. (2014), or are neural models trained on annotated data which is available in sufficient quantities only for high resource languages (Jha et al., 2018). Building morphology analyzers for low-resource languages remains a challenging task. For low-resource languages, morphological resources are sparse or virtually nonexistent. Multilingual models have shown promising results for cross-lingual transfer from high-resource to low-resource languages (Wu

*Work done while at IIT Bombay

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 14–26

Percentage of data points with a particular feature marking is present							
Gender	Number	Person	Tense	Aspect	Case	Modality	Others
60.4	94.8	82.1	58.5	35.5	11.0	11.0	27.5

Table 1: Combined statistics of annotated data (across languages) available for various tags. We work with gender, number, and person as they have the highest proportion as compared to other features and are common to noun and verb morphology. We don't use tense as it is not relevant to nouns. More details in section 3

and Dredze, 2019; Lauscher et al., 2020). The main goal of our work is to increase NLP inclusivity. The primary obstacle one encounters while expanding the coverage of NLP models is the lack of usable (annotated) data for most languages; collecting (annotated) data is a painstaking task, especially for endangered languages. When data is sparse, we turn to linguistics to help exploit universalities across languages.

In this work, we study the multilingual capability of mT5 (Xue et al., 2021) to carry out cross-lingual transfer of morphological features and extract the root words given the surface forms. We also test the multilinguality hypothesis that, in the presence of annotated examples of source languages, the required number of annotated examples of the target language to get identical results reduces. We carry out this analysis of cross-lingual transfer within language families and across (language) families and provide pointers to effective usage of multilingual data. The languages we carry out morphological analysis are of the Dravidian family (Tamil, Telugu, and Kannada) and the Indo-Aryan family (Bengali, Hindi, and Marathi). We also give a brief account of the state of datasets available for morphological analysis and their challenges. We finetune mT5 for gender, number, and person tagging for verbs and nouns in six Indian languages: *Marathi, Hindi, Bengali, Tamil, Telugu, and Kannada*. The features: gender, number, and person (GNP) are hereby referred to as morphosyntactic description (MSD) tags. The current state of the datasets and inconsistency of annotation across languages limits our analysis to GNP tags of verbs and nouns.

Our contributions are as follows:

- We test the multilinguality hypothesis that the availability of annotated data of source languages reduces the number of examples of target language required to outperform the monolingual baseline.
- We study inter-family and intra-family transfer in the context of GNP tagging and root word extraction for languages from Dravidian and Indo-Aryan families.
- We analyze how multilingualism helps in the morphological analysis of verbs and nouns, root word extraction, and test the model's ability to generalize to unseen suffixes.

2 Related Work

Morphological analysis: For morphological analysis, SIGMORPHON (Nicolai et al., 2021) has been one of the venues organizing shared tasks and workshops related to computational morphology and multilingual morphological analysis, especially in the low resource scenarios. Shared tasks such as Cotterell et al. (2016, 2017, 2018), etc. looked at morphological reinflection with an increasing number of languages each year. For morphological reinflection, the output is the surface form, and the inputs are: a root word (or any other form of the root word) and desired features in the surface form (the output). Task 2 in Cotterell et al. (2018) as well as McCarthy et al. (2019) explored morphological analysis and reinfection in context. Jin et al. (2020) and Wiemerslage et al. (2021) were aimed at unsupervised clustering of paradigms, wherein given a lemma list, the goal is to output all the possible forms of a lemma. Morphosyntactic lexicon generation is one task closely related to morphological analysis; Faruqui et al. (2016) used graph-based semi-supervised learning for label propagation. Hulden et al. (2014) used a semi-supervised approach for lexicon construction from concrete inflection tables by generalizing the inflection paradigms from the tables provided. For morphology resources, apart from UniMorph (Batsuren et al., 2022; McCarthy et al., 2020), the MorphyNet database (Batsuren et al., 2021) is a large dataset of methodologically annotated surface forms spanning 15 languages and is extracted from Wiktionary. There have also been efforts to create task-specific models for various components of cross-lingual morphological tagging (Cotterell and Heigold, 2017a; Malaviya et al., 2018)

Indian language morphology: Regarding resources for Indian languages, Arora et al. (2022) points out resource scatteredness (rather than

scarcity) as the primary obstacle to developing South Asian language technology and proposes the study of language history and contact as one of the potential solutions. Workshops like DravidianLangTech (Chakravarthi et al., 2021) and WILDRE (Jha et al., 2020) are dedicated specifically to the development of technologies and resources for Indian languages. The UniMorph database (McCarthy et al., 2020) has been one of the recent efforts to extend the coverage of computational morphological resources. Cotterell and Heigold (2017b) trained bidirectional character-based LSTM-based models to demonstrate the effectiveness of the cross-lingual transfer. They have trained bilingual models for languages from Romance, Slavic Germanic, and Uralic families. Gupta et al. (2020) trained various sequence labelling models for Sanskrit. Nguyen et al. (2021) trained transformer-based models for various NLP tasks such as PoS tagging, Morphological feature tagging, and dependency parsing for over 100 languages. Nair et al. (2021) carried out a comparative study of existing morphological analyzers for Indian languages to conclude that although morphological analyzers exist for Indian languages like Sanskrit and Malayalam, they are not accurate as compared to the high resource baselines. Elsner (2021) probed an analogical memory-based framework for one-shot morphological transfer to study the abstract representational concepts learned by the transfer networks.

3 Dataset Challenges

Creating a multilingual morphology analyzer would require a union of the sets of features across all the languages and all the parts of speech. The morphological features are modeled as categorical variables in fixed output space. The modeling difficulties arise primarily due to the following: (1) absence of feature annotations for Indian languages, (2) lack of data for all the parts of speech (PoS) except verbs and nouns and (3) variance of markings across PoS and languages. The dataset only contains data for verbs and nouns, which restricts our analysis to those PoS. For these PoS, the feature data is primarily available for Gender, number, and person compared to other features, so we carry out transfer analysis for only those features. We provide a summary of annotated data available in Table 1. Gender, number, and person also happen to be morphological features that are common to nouns and verbs. We provide detailed statistics of the UniMorph dataset in appendix A.

We have used various data sources to demonstrate the scalability of the morphology analyzer to 6 Indian languages. For languages Hindi, Telugu, Kannada, and Bengali, we have used the UniMorph 3.0 (McCarthy et al., 2020) dataset. The number of examples varies across languages. For Bengali, the number of examples available is 4443; for Kannada, it is around 6400; for Hindi, there are about 54K examples, while Telugu has about 1500 examples. All the examples in the UniMorph dataset are either verbs or nouns. For Tamil, morphologically annotated data from the Tamil dependency treebank (Ramasamy and Žabokrtský, 2014) was used. The number of annotated words (verbs+nouns) in the tree bank is 9521, all of which were used. For Marathi, we used the dataset from Bapat et al. (2010). The dataset consists of around 21k annotated words, out of which we used 15k words, nouns, or verbs, to have consistency with other datasets. Although there are other sources of data, such as Bhat et al. (2017), we stick to the UniMorph dataset wherever possible to ensure higher annotation accuracy. The scope of our work limits demonstrating the usefulness of cross-lingual transfer for morphological analysis, so dataset selection and optimizing the number of examples for creating the best morphology analyzer remains a challenge for future research.

4 Modelling Details

4.1 Morphological analysis as text to text problem

The Multilingual T5 (Xue et al., 2021) is a massively multilingual pre-trained text-to-text transformer model released by Google in 2020. It is pre-trained on the Common Crawl-based dataset and covers 101 languages. It is an encoder-decoder sequence generation model, unlike mBERT, which is an encoder-only multilingual model. Our task of root word extraction requires the generation of text sequences, so we use an encoder-decoder model to avoid training a decoder separately for the given languages. We use the mT5 base model with 580 million parameters for our experiments.

As mT5 is a text-to-text sequence generation model, the tags are generated as a sequence of text, one after the other. The input to the model is the surface form of the words, and the model generates the gender, number, and person tags as a text sequence. Not all the words in the dataset would be

Modeling Strategy	Accuracies For Marathi			
	Monolingual		Multilingual	
	Root Word	MSD Tagging	Root Word	MSD Tagging
Joint model	42.2	79.7	53.2	84.6
Multitask model	26.2	**86.5**	52.2	88.2
Independent Model	**78.2**	81.2	**86.4**	**95.2**

Table 2: Comparing three modeling strategies for root word extraction and MSD tagging. Training a separate multilingual model for both tasks is the best-performing strategy. We provide details in section 4.2

morphologically marked for GNP; for example, in the case of person marking for nouns, the markings are only present on the pronouns (and the surface form changes according to the person). In contrast, the surface form remains the same for common and proper nouns, irrespective of the person. In such cases, where the marking is either trivial or marking is not present on the word or where the marking cannot be inferred from the surface form itself, the model's expected output is the tag 'unknown'. The datasets we use contain morphological tags without context; we, therefore, predict the tags solely based on the markings present on the words rather than the context and assign the tag 'unknown' to the words for which tags cannot be predicted without context. For all the experiments, unless and otherwise stated: we use the following evaluation strategy: We firstly remove the 20% data (randomly sampled) for each language (which is used for evaluation) and use the remaining 80% data for experiments. We ensure that the randomly sampled data contains unseen paradigms; no surface form of the lemma is present in the training dataset. Across the monolingual and multilingual experiments, the evaluation data remains the same. To avoid the error variation due to bias in sampling (wherein the test set contains all the paradigms available in the training set), we use k-fold cross-validation (with k=5) and report average numbers. The epochs used were 7-15 based on performance on validation data. As far as metrics for measuring model performance are concerned, we report per-tag accuracy for each of the GNP tags, and overall accuracy. The overall accuracy denotes the percentage of instances for which all three tags are predicted correctly by the model. For root word extraction, we consider exact-string match based accuracy.

4.2 Three modelling strategies

We consider three modelling strategies for MSD tagging and root word extraction.

- **Joint model**: We first use the mT5 as a sequence prediction model wherein the input is the surface form, and the outputs are the root words and MSD tags. The root word and MSD tags are generated as a sequence, with the root word being generated first, followed by MSD tags: gender, number, and person (in that order).
- **Multitask model**: In the second setting, we use mT5 as multitask model, with MSD tagging and root word extraction being treated as two separate tasks. We prepend a prefix (string) to the input to specify which task should be performed.
- **Independent model**: In the third setting, we train separate models for root word extraction and MSD tagging, with MSD tags being predicted as a sequence of letters and the input being the surface form.

The input to the model for the second task is the surface form, along with a prefix specifying the task. It should be noted that the choice of prefixes is arbitrary, as long as they are different for each task. While fine-tuning, we add explicit language flags with the respective surface words.

We compare the training strategies in Table 2. The joint sequential prediction leads to the least accuracy in both tasks. Although the multitask framework has higher accuracy than the joint prediction for MSD tagging, it has the lowest accuracy for root word prediction. The multitask framework is expected to have high accuracies because both tasks (MSD tagging and root word extraction) are closely related to each other in the following way: The suffix determines the MSD tags of the surface form, and thus identifying the suffix is an important part of MSD tagging while stripping away the suffix is one of the aspects of root word extraction. The joint multitask training leads to the mixing of outputs (the outputs of both the tasks are in different languages: The MSD tags are in English while

Language	Monolingual accuracies				Multilingual accuracies			
	Gender	Number	Person	Overall	Gender	Number	Person	Overall
Tamil	80.1	87.9	86.3	79.3	86.3	91.7	89.7	**85.4**
Telugu	78.9	97.7	87.4	76.2	78.6	98.3	87.6	**76.5**
Kannada	84.0	88.1	82.6	70.1	87.3	95.7	86.8	**81.7**
Marathi	88.2	87.2	89.3	90.2	96.7	95.9	97.7	**95.6**
Hindi	92.1	85.1	56.9	**53.5**	99.0	89.1	58.3	52.6
Bengali	99.3	94.3	85.0	85.8	99.2	98.3	90.8	**90.4**

Table 3: Demonstrating the benefit of multilingual models over monolingual models for all three tags. The per-tag accuracies and overall accuracies show an increase for all languages except Hindi and Telugu, which show a slight decrease in overall accuracy (but the per-tag accuracy increases for all languages). We provide details of the experiments in section 4.1

the root words are in the same language as the surface form), as observed during the performance on the test set. Training a separate model for both tasks yields the highest performance, and we use the strategy for all our subsequent experiments.

5 Low Resource Morphological Analysis Experiments

5.1 Multilinguality hypothesis

We test the multilinguality hypothesis by comparing monolingual models with multilingual models. As seen in Table 3, which shows per-tag accuracy for each gender, number, person tag, along with overall accuracy, multilingual models outperform the monolingual models for most of the languages except for Hindi and Telugu. One of the reasons for worse performance of multilingual model for Hindi is that the Hindi data containers phrases and post-positions with GNP markings, which are not present in other languages. Thus, adding multilingual data leads to drop in model performance due to confusion between word-based markers and post-position based markers. For Dravidian languages, the overall increase in the accuracy of the multilingual model is negligible in the case of Telugu (as compared to the monolingual baseline), the other two languages, Tamil and Kannada, show around 7.8% increase in overall accuracy. Multilingual models also show better scores in the case of per-tag accuracies for all the Dravidian languages, with gender tag having the highest average increase of 4.03% (averaged over languages).

As also seen in column 2 of Table 2, there is an increase in the accuracy of root word extraction and MSD tagging for Marathi. We show more evidence of the multilinguality hypothesis for MSD tagging through controlled experiments on Bengali and Kannada. We chose these two languages to study the transfer because (1) Kannada shows the highest increase in overall accuracy among the Dravidian languages, and (2) the number of annotated examples of Bengali is the least among the Indo-Aryan languages. Choosing these two languages helps us clearly observe the effect of cross-lingual transfer and the low resource scenario. Tables 4 and 5 show that the multilingual models outperform the monolingual models irrespective of the source languages, with the increase in accuracy being the highest (around 54% for Bengali and 33% for Kannada) in sparse data scenario, where the number of examples of the target language is 1000. Tables 6 and 7 show evidence of the multilinguality hypothesis for root word extraction.

5.2 Inter-family and intra-family transfer

To study cross-family and intra-family transfer, we use Bengali and Kannada. Bengali has the least number of examples in the Indo-Aryan family and shows the highest increase in accuracy with the addition of multilingual data. Kannada shows the highest increase in overall accuracy when going from a monolingual to a multilingual setting. We do this by varying the number of examples of Bengali in the train set to simulate the low-resource scenario. We also add various sets of languages as a source to check inter-family and intra-family transfer. Note that the last row in all the tables named 'All Languages' implies that the data of all six languages were used for training. We study the effectiveness of (family-based) multilingual data by analyzing inter-family and intra-family transfer. In the case of Bengali, we observe that intra-family transfer from languages of the Indo-Aryan family,

viz., Marathi and Hindi, lead to, on average, 2.82% more accuracy as compared to transfer from the Dravidian family for MSD tagging (Table 4). For Kannada, the increase in accuracy from monolingual baselines is more from the languages of the Dravidian family as compared to the Indo-Aryan family when the number of examples of Kannada in the training data is 1000 (Table 5). In all other cases, the increase in accuracy with Dravidian languages is either less or similar to that with Indo-Aryan languages as a source. When languages from both families are used as source languages, we observe a sharp increase in accuracy for the root word extraction in Bengali and Kannada. For both the languages, Bengali and Kannada, there is a decrease in accuracy when all the languages are used as source languages, compared to the setting where languages from a particular family are used as source languages.

6 Analysis

In this section, we provide further analysis of the cross-lingual transfer of MSD tags for verbs and nouns and root word extraction.

6.1 GNP tagging for verbs and nouns

In Table 3, we note that the increase in overall accuracy in the case of the multilingual model is the highest for Kannada in the Dravidian family as compared to the monolingual model. Bengali has the least number of annotated examples and shows the highest increase in accuracy from monolingual baseline in the Indo-Aryan family. We dive further into the accuracies of Kannada and Bengali. To investigate the sources of multilingual signals, we conduct experiments separately for nouns and verbs.

Nouns: For nouns, the person feature is trivially third (except for pronouns), and the number feature can be inferred from the suffix, but the gender assignment is arbitrary, and we may require a dictionary to get the gender of the nouns. So, if the nouns (present in the test set) have not been seen during training by the model, one of the potential sources of signal regarding gender is the multilingual data. Another source of signals for gender is also the context that the model has seen during the pretraining (for example, the gender of the nouns is marked on verbs). It is hoped that the gender signals will be captured in the representations learned during the multilingual pretraining. The shared latent space, learned by the multilingual models, is assumed to cluster the words of the same meaning in different languages close to each other.

To test the hypothesis regarding the gender of nouns, we test the accuracy of Kannada and Bengali nouns with various training data from multiple languages. As the gender signal can be dictionary-based, we see that the accuracy increases irrespective of the source languages, as shown in Table 9 and Table 12 in appendix B. For both the languages, Bengali and Kannada, we note that the gender accuracy is higher when the source languages are Marathi and Hindi. The higher accuracy is because the number of training examples of Hindi and Marathi combined is around 70k, while the number of examples of all Dravidian languages combined is about 17K, so more the number of nouns in the training set, more would be the hope of getting dictionary signals. As additional evidence, we also carry out zero-shot transfer for nouns of each language. The training data consists of nouns from all the available languages, and the test data contains nouns from the target language, as shown in Table 13 in the appendix B. The zero-shot gender predic-

Source Languages	Number of Bengali training examples		
	1000	2000	3000
Monolingual	30.8	82.2	85.8
Marathi, Hindi	**85.5**	**89.4**	**90.4**
Tamil, Telugu, Kannada	83.1	87.6	86.1
All languages	73.8	88.9	89.8

Table 4: Bengali MSD tagging accuracies demonstrating effectiveness of intrafamily transfer and multilinguality over monolingual model for low resource setting. More details in section 5.2

Source Languages	Number of Kannada training examples			
	1000	2000	3000	4000
Monolingual	33.2	52.8	65.1	81.7
Marathi, Hindi, Bengali	63.9	77.2	81.4	84.9
Tamil, Telugu,	69.4	74.8	**82.8**	**85.4**
All languages	**69.8**	**76.3**	78.3	82.2

Table 5: Kannada MSD tagging accuracies demonstrating effectiveness of intra-family transfer and multilinguality over monolingual model for low resource setting. More details in section 5.2

Source Languages	Number of Kannada training examples			
	1000	2000	3000	4000
Monolingual	23.2	31.2	40.9	51.2
Marathi, Hindi, Bengali	67.2	70.8	76.7	80.2
Tamil, Telugu,	69.1	71.5	72.9	77.5
All languages	**70.4**	**72.6**	**78.8**	**83.2**

Table 6: Kannada root word extraction accuracies demonstrating multilinguality hypothesis. More details in section 5.1

tion accuracy is non-trivially high for all languages except Tamil (as compared to the case where only verbs are used as source data, wherein we get trivial test accuracies). Tamil has less accuracy for gender as compared to other languages because the number of genders in the Tamil dataset is five, and in a zero-shot setting, the model has no way of knowing the presence of five genders.

Verbs: In the case of verbs, all the features: gender, number, and person can be inferred from the suffix. Our hypothesis here is that increase in the accuracy of verbs in the multilingual setting depends on the source language data available for training. As seen in Table 8, the highest increase in the accuracy of Bengali verbs is seen when the source languages are from the same family. The gender accuracy is almost the same for all the languages as Bengali is a gender-less language, and there are no markings of gender on verbs. In the case of Kannada, as shown in Table 11, the highest increase is observed when the source language is Tamil and Marathi. A Significant increase in accuracy when source data from Marathi is used provides evidence of historical contact between these two languages, as has been discussed in Sengupta and Saha (2015).

Source Languages	Number of Bengali training examples		
	1000	2000	3000
Monolingual	32.2	51.2	74.8
Marathi, Hindi	85.2	92.3	95.2
Tamil, Telugu, Kannada	84.6	91.2	93.3
All languages	**90.1**	**92.8**	**96.9**

Table 7: Bengali root word extraction accuracies demonstrating positive transfer from various subsets of source languages. More details in section 5.1

Source language	Accuracy for Bengali Verbs			
	Gen	Num	Per	Overall
Monolingual	99.3	94.2	84.6	76.2
Marathi	99.2	92.6	89.4	88.9
Hindi	99.2	**93.2**	**90.7**	**90.0**
Tamil	99.6	92.6	86.8	86.8
Telugu	99.1	91.6	84.2	83.9
Kannada	99.2	91.3	88.2	87.3
Hindi, Marathi	99.8	**93.4**	**90.5**	84.1
Tamil, Telugu, Kannada	99.8	91.6	89.8	85.8

Table 8: Analysis of Bengali Verbs demonstrating transfer from various families and languages. More discussions in section 11

Training languages	Accuracy on Kannada Nouns		
	Gender	Number	Overall
Monolingual	96.0	90.4	82.9
Tamil, Telugu	94.0	94.9	89.4
Marathi, Hindi	**96.9**	97.3	94.4
All Languages	96.0	**97.7**	**95.7**

Table 9: Testing cross-lingual transfer for Gender and Number tags in the case of Kannada Nouns

The historical contact also shows the reason behind the highest increase in overall accuracy when the source languages are Marathi and Hindi (Table 9).

Generalization: We also test the model's generalization ability to unseen patterns. For example, the suffix 'raha hei' in Hindi represents masculine, third person, and singular. We remove all instances of the suffix from the train set, add them to the test set, and check the accuracy of the model on it in multilingual and monolingual settings. In the case of monolingual and multilingual settings, the model's overall accuracy is 50% for GNP tagging; the tags gender and number are correctly predicted for all the test instances, while the person tag is correctly predicted for 50% of all the instances. The number and gender can be inferred from the suffix itself; however, the person tag depends on the verb as well as the context, thus leading to confusion for the model (as we are not using the context currently.)

Number of training examples	Bengali		Kannada	
	(1) Input same as outputs	(2) Surface forms and roots	(1) Input same as outputs	(2) Surface forms and roots
zero-shot	12.2	18.2	8.6	12.1
1000	90.3	90.1	72.2	70.4
2000	94.8	92.2	71.2	72.6
3000	**97.9**	96.9	73.9	78.8
4000	-	-	74.4	**83.2**

Table 10: Role of copy bias in root word extraction. Adding inputs same as outputs for source languages has results comparable to the case when inputs are surface form and outputs are root words. (Note: Number of available training examples of Bengali is 3000) More details in section 6.2

Source language	Accuracy for Kannada Verbs			
	Gen	Num	Per	Overall
Monolingual	83.0	95.8	82.7	73.1
Marathi	87.6	**96.3**	83.3	76.0
Hindi	85.6	95.5	90.3	81.9
Tamil	88.0	96.7	92.3	**84.1**
Telugu	80.6	94.6	74.1	66.1
Bengali	84.6	97.4	91.8	81.8
Hindi, Marathi, Bengali	81.4	94.4	83.9	76.0
Tamil, Telugu	**87.7**	97.2	91.6	83.5

Table 11: Analysis of Kannada Verbs demonstrating transfer from related families and languages. More discussions in section

6.2 Root word extraction

To test cross-lingual transfer in the case of root word extraction, we test the copy bias learned by the model. The copy bias is an essential part of the learning process for root word extraction, as the output contains most of the characters present in the input except for a suffix. As can be seen in Tables 7 and 6, the root word extraction accuracy increases to a similar extent, irrespective of the source language. We test the copy bias by adding training examples from source languages such that the input and output are the same. The comparison of the effect of copy bias with our standard setup where the source inputs are surface form and source outputs are root words is shown in Table 10. The table highlights that copy bias plays a role in root word extraction and cross-lingual transfer of morphological knowledge (such as the similarity between morphemes) across the shared embedding space is limited.

7 Conclusions

In this paper, we tested the multilinguality hypothesis for root word extraction and morphosyntactic descriptors (MSD) tagging. We trained multilingual models for MSD tagging and root word extraction using data of six Indian languages spanning two families of the Indian subcontinent. We demonstrated the effectiveness of data from languages of the same and different families and how it can be leveraged to train morphological analysis models for low resource languages. We also analyzed how cross-lingual transfer of morphological knowledge happens for nouns and verbs along with the copy bias, which forms a significant component of the root word extraction. Our framework can be extended to multiple tags as well as more low resources languages as annotated data becomes available. We see our work as an important step in the direction of bias-aware machine translation to morphologically rich languages.

8 Limitations

One of the limitations of our work is the unavailability of context data and unavailability of phrase-based annotations for all languages except Hindi. The unavailability of phrase-based annotations prevents the usage of universal tags because markings that are present on a single word in highly agglutinative languages like Marathi or Tamil get expressed on 2–3 words in isolating or fusional languages like Hindi or Bengali (where markings are present on post-positions). The benefits of using phrase level morphology over token level morphology have been discussed in Goldman and Tsarfaty

(2021). For example, the word 'sochega' in Hindi will have MSD tags: future tense and male gender, while in English, it would take two words, 'he will think' to express the same amount of morphological information. The presence of contextual data can also help to disambiguate MSD tags. The other limitation of our work is the mismatch between the languages for which pretrained models (especially encoder-decoder models) are available and the languages for which we have the annotated data. For example, UniMorph dataset contains annotated examples for Assamese and Sanskrit, but we do not have multilingual pretrained encoder-decoder models for these languages.

9 Ethics Considerations\Broader Impact

Our work is on morphological analysis of low resource languages. We aim to increase the coverage of NLP tools through our work. It is inline with making language technologies accessible for wider range of audiences who do-not have commonly researched high resource languages like English, French as their native language. Our work is also a step towards automating the process of documentation of endangered languages.

Acknowledgements

We thank Simran Khanuja for participating in the early phases of this research. We also thank CFILT Lab, IIT Bombay for providing resources and some of the datasets. We thank Kuzman Ganchev and Srini Narayanan for comments on the draft. We would like to thank CMiNDS department, IIT Bombay along with Google India Research Lab for providing with the opportunity for conducting this research. Finally, we would like to thank Matthias Bauer for providing help with writing and structuring of the final draft.

References

Ankita Agarwal, Pramila, Shashi Singh, Ajai Kumar, and Hemant Darbari. 2014. Morphological analyser for hindi – a rule based implementation. *International Journal of Advanced Computer Research*, 4.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Mugdha Bapat, Harshada Gune, and Pushpak Bhattacharyya. 2010. A paradigm-based finite state morphological analyzer for Marathi. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, pages 26–34, Beijing, China. Coling 2010 Organizing Committee.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, pages 39–48.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. 2017. The hindi/urdu treebank project. In *Handbook of linguistic annotation*, pages 659–697. Springer.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar M, Parameswari Krishnamurthy, and

Elizabeth Sherly, editors. 2021. *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv.

Ryan Cotterell and Georg Heigold. 2017a. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759.

Ryan Cotterell and Georg Heigold. 2017b. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Micha Elsner. 2021. What transfers in morphological inflection? experiments with analogical models. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 154–166, Online. Association for Computational Linguistics.

Manaal Faruqui, Ryan McDonald, and Radu Soricut. 2016. Morpho-syntactic lexicon generation using graph-based semi-supervised learning. *Transactions of the Association for Computational Linguistics*, 4:1–16.

Omer Goldman and Reut Tsarfaty. 2021. Well-defined morphology is sentence-level morphology. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 248–250, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashim Gupta, Amrith Krishna, Pawan Goyal, and Oliver Hellwig. 2020. Evaluating neural morphological taggers for sanskrit. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 198–203.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

Girish Nath Jha, Kalika Bali, Sobha L., S. S. Agrawal, and Atul Kr. Ojha, editors. 2020. *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*. European Language Resources Association (ELRA), Marseille, France.

Saurav Jha, Akhilesh Sudhakar, and Anil Kumar Singh. 2018. Multi task deep morphological analyzer: Context aware joint morphological tagging and lemma prediction. *ArXiv*, abs/1811.08619.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena

Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Jayashree Nair, L. S. Aiswarya, and P. R. Sruthy. 2021. A study on morphological analyser for indian languages: A literature perspective. In *Advances in Computing and Data Sciences*, pages 112–123, Cham. Springer International Publishing.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Garrett Nicolai, Kyle Gorman, and Ryan Cotterell, editors. 2021. *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Online.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology Matters: A Multilingual Language Modeling Analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2014. Tamil dependency treebank v0.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Wolfgang Seeker and Özlem Çetinoğlu. 2015. A Graph-based Lattice Dependency Parser for Joint Morphological Segmentation and Syntactic Analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Debapriya Sengupta and Goutam Saha. 2015. Study on similarity among indian languages using language verification framework. *Advances in Artificial Intelligence*, 2015:1–24.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. 2017. Modeling target-side inflection in neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, Copenhagen, Denmark. Association for Computational Linguistics.

Adam Wiemerslage, Arya D McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. Findings of the sigmorphon 2021 shared task on unsupervised morphological paradigm clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–81.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

A Appendix: Satistics of UniMorph Dataset

The UnimOrph dataset's statistics are shown in table 14. A total of 26 features are available in the meta-data of the UniMorph dataset. They include Aktionsart, Animacy, Argument marking, Aspect, Case, Comparison, Definiteness, Deixis, Evidentiality, Finiteness, Gender, Information Structure, Interrogativity, Language Specific features, Mood, Number, Other, Part of speech, Person, Polarity, Politeness, Possession, Switch reference, Tense, Valency, Voice. For most Indic languages, the annotations are present for not more than eight features

per language. The set of features for which annotations are current varies across languages. We give the proportion of words in the dataset for which feature annotations are present. We provide statistics for Gender, Number, Person, Tense, Aspect, and Modality, characteristic features of verbal morphology. We also provide statistics for case, number, number, and person for nouns. The 'others' section represents the features with the highest proportion of tags, from gender, number, person, tense, aspect, modality, and case. Also, one thing that must be noted is that the amount of data available for verbs is almost 5 times the data available for nouns for most of the languages, so the number in the 'total' row is dominated by statistics of verb. For Hindi, the nouns data is completely absent.

B Appendix: Cross-Lingual transfer Nouns—Additional Tables

Training languages	Accuracy on Bengali Nouns		
	Gen	Num	Overall
Monolingual	96.81	79.62	76.85
Tamil, Telugu, Kannada	95.18	91.66	85.18
Marathi, Hindi	98.21	92.23	87.37
All	**98.45**	**92.7**	**90.7**

Table 12: Testing cross-lingual transfer for Gender and Number tags in the case of Bengali Nouns

Target language	Zero Shot Test accuracy for nouns		
	Gender	Number	Overall
Marathi	68.2	76.4	66.4
Telugu	69.6	59.7	48.1
Bengali	55.1	65.5	50.2
Kannada	56.2	61.2	47.3
Tamil	15.1	67.1	13.2

Table 13: Zero-shot accuracies for gender and number tagging of nouns showing the help of multilingual signals for gender. More details in section 6.1

Lang	POS for which data is available	Percentage of data points with a particular feature marking is present							
		Gen	Num	Per	Ten	Aspect	Case	Modality	Others
Hindi	Verbs	94.7	99.0	95.2	34.1	89.1	0	27.0	35.2
	Nouns	-	-	-	-	-	-	-	-
	Total	94.7	99.0	95.2	34.1	89.1	0	27.0	35.2
Bengali	Verbs	-	100	86.9	86.9	60.8	-	2.1	52.1
	Nouns	66.6	-	-	-	-	80.8	-	19.8
	Total	8.0	88.9	75.6	75.6	52.9	10.5	1.8	45.3
Kannada	Verbs	46.6	100	89.2	46.2	-	-	19.6	20.7
	Nouns	-	100	-	-	-	100	-	-
	Total	36.6	91.4	70.5	37.1	0	20.9	15.5	16.8
Telugu	Verbs	50.0	100	100	100	-	-	-	13.7
	Nouns	-	100	-	-	-	100	-	-
	Total	43.7	100	87.2	87.2	0	12.7	0	11.3
Combined	Verbs	47.8	99.7	92.8	77.5	37.4	-	12.1	30.4
	Nouns	16.6	50	-	-	-	70.2	-	4.9

Table 14: Statistics of UniMorph dataset