Neural Morphology Analysis - A Survey

Siddhesh Pawar[†] and **Pushpak Bhattacharyya**[†]
[†]IIT Bombay
{17d170011, pb}@iitb.ac.in

Abstract

Morphology analysis is a foundational NLP task with relevance in language documentation, machine translation, language modeling, etc. Building morphological analyzers is a resource intensive process limiting their availability to a few high web-resource languages. Neural models such as RNNs, transformers, etc have been applied recently to morphology analysis tasks. These methods have shown great improvements over the baselines, especially in a resource scarce scenario. In this survey, we focus on supervised and unsupervised approaches that have been tried for morphology analysis. We also provide a survey of the current state of resources that are available for morphology analysis.

1 Introduction

The term morphology refers to the phenomenon of word formation in a rule-governed way. Morphology analysis is the first step of processing in the NLP pipeline. The focus is on analyzing the internal structure of words, understanding the meaning and function related to each part (also called morpheme), and figuring out how different morphemes can be combined to make valid words. Even in the transformer era, wherein the entire NLP pipeline is replaced with a transformer, 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) as statistical subword tokenization cannot capture morphological alternations and non-concatenative morphology. A morphology analyzer can also help in speeding 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.

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 morphologically annotated data which is available in sufficient quantities only for high resource languages (Jha et al., 2018). Building morphology analyzers for low-resource languages still 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 languages to low-resource languages (Wu and Dredze, 2019; Lauscher et al., 2020). The shared representations learned by these models allow for effective few shot learning, thus facilitating coverage to low-resource languages.

Computational morphology consists of varied tasks such as root word extraction, surface word segmentation, MSD tagging, learning of allomorphy and morphophonological variations, morphological (re)inflection, paradigm completion, paradigm clustering, simulating cognitive processing of morphology. Our focus in this paper is primarily on morphology analysis and morphology generation. For an overview of other tasks, we refer the reader to Liu (2021). Morphology analysis consists of two sub-tasks: root word extraction and MSD feature tagging. In the case of morphology generation, the inputs are the MSD and the root word and the output is the surface form with features as described in the input. We provide an example highlighting distinction between these tasks in figure 1. We survey unsupervised approaches to morphology analysis in section 2. We then look at

supervised approaches in section 3. We then provide an overview of available resources for South Asian Languages in section 5. We finally review the shared tasks in SIGMORPHON in section 7.

2 Unsupervised Approaches to Morphology Analysis

The key idea in the unsupervised setting is to design an algorithm that takes a text corpora (which is unannotated) as input and provides an analysis of the morphological structures that are present in the language. The main motivation to study unsupervised approaches is rooted in the quest to find out the languages learning phenomenon that goes on inside the human brain. (Creutz and Lagus, 2004) propose a Morfessor baseline member that uses Maximum-Likelihood (ML) estimate. Expectation Maximisation is used to optimise the model (EM). At first, the words are divided at random. Words are separated iteratively by drawing morpheme lengths from a Poisson distribution. Splits are either accepted or rejected according to the rejection criteria which has two conditions; rare morphemes and one letter morphemes are rejected. To produce a power-law distribution, (Goldwater et al., 2005) offer a two-stage model in which words are first generated by a generator component and then the frequencies of the words are estimated by an adapter. The adaptor runs a Pitman-Yor process by locating the words in tables in a rich-get-richer fashion. Instead of inducing the morphemes in each language separately, (Snyder and Barzilay, 2008) construct a non-parametric Bayesian model that uses bilingual parallel corpora to induce commonly occurring morphemes (abstract morphemes) inside parallel short phrases. It is a hierarchical Bayesian model where the defined distributions are drawn from Dirichlet processes. Although it has only been tested on bilingual corpora, the model can also be extended to induce morphemes across multiple languages. Authors in (Chan, 2006) propose a Latent Dirichlet Allocation (LDA)-based method, which is also a generative probabilistic model, in which data sets are created using a three-level hierarchical Bayesian mode. When it is applied to topic modelling, the three levels consist of documents, topics and a vocabulary. (Chan, 2006) applies a similar approach by replacing the documents, topics and a vocabulary with the suffixes, stems and paradigms, where the latent classes are the paradigms to be induced.

2.1 Geometric Approaches

The geometric approaches to morphology analysis deal with travelling and searching effectively in a geometric space. The the process of specifying a morphological rule can be thought of as locating it as a point in a space of very high dimension, and the task of finding the correct grammar (rule) can be thought of as traveling through that space. Three main approaches that have been explored in this context are: Minimum Description Length (MDL), Gibbs sampling, and adaptor grammars. These models have been built on the concept of probabilistic models and involve finding the rules (paths in the morphological space) that maximize a probabilistic score. We describe the approaches in detail below:

2.1.1 Minimum Description Length (MDL)

Minimum description length was proposed by (Goldsmith, 2006) in the context of unsupervised morphology leaning. The main idea is to quantify the information contained in particular morphological description of a particular set of data D as sum of two quantities: the complexity of the overall grammar G used to provide the description, and the number of bits needed to represent the data D , given G, a probabilistic grammar.The first term represents the algorithmic complexity of the algorithm while the second term represents the goodness of fit of particular analysis (that is the set of rule) of dataset given the grammar (This can be thought of as the quantity of information in the corpus that is not explained by the grammar). The minimum description length is based on the principles that: Every regularity in data may be used to compress that data and Learning can be equated with finding regularities in data. Goldsmith introduces the morphological structures called 'signatures' to encode the data. A signature represents the inner structure of a list of words that have similar inflective morphology. The morphology of a corpus is represented in three lists: an affix list, a stem list, and a signature list. The affix and the stem list contain the letters, whereas the signature list only contains pointers to stems and affixes. The aim is to find the morphology that will analyse the corpus in its most compact state. Morfessor (Creutz and Lagus, 2004) is another state of the art model in the field. It engages the MDL principle to minimise the length of a codebook (A code-

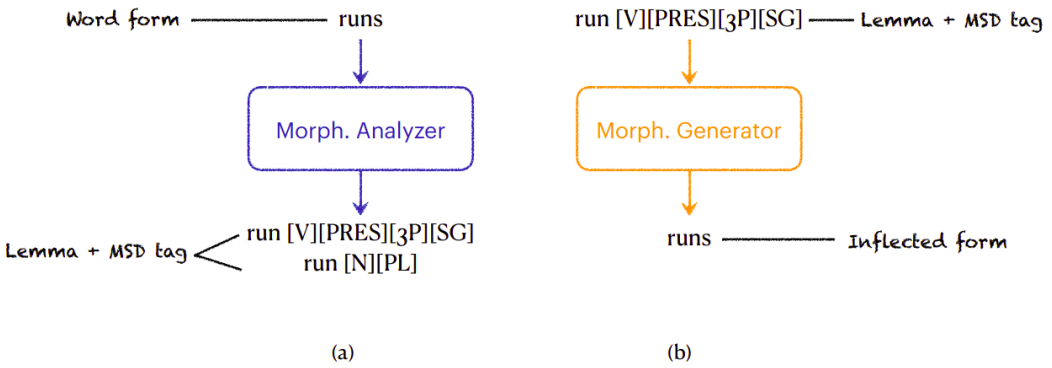

Figure 1: (a) Morphology analysis consists of lemmatization and morphological tagging on the English word 'runs'. (b) is a morphological generator produces the inflected form corresponding to the English lemma run and the MSD tag V;PRES;3P;SG. Source: (Liu, 2021)

book consists of the morphemes that will generate a corpus.) and a corpus. The length of a corpus is computed using the maximum likelihoods of the morphemes, whereas the length of a codebook is the summation over all morphemes' lengths. Morfessor Baseline deploys a recursive segmentation where each discovered morpheme is analysed recursively as long as it improves the cost.

3 Supervised Approaches to Morphology Analysis

3.1 Morphology Analysis as Learning task

Morphology analysis can be thought of as a learning task in three different ways. One way of looking at is as a classification task wherein the model takes each character as input, builds a representation of the entire word and then outputs one class (or multiple classes like MSD tags) as the output.This can be regarded as many to one case. In the many-to-many case, the model first reads in the characters in the text, generates the representation of the text and then outputs another sequence of symbols which may or may not be of the same length as the input. This is known as sequence to sequence transduction. The task of morphological inflection and reinfection falls in this category. Another alternative way to look at the task of morphological analysis is sequence labelling. In this case, the model takes a sequence of character and labels each and every character (the labelling is done on the basis of actions that need to be taken on that particular character). For example while obtaining the root word of the word going, the output would be label for each charter pertaining to deletion or retention of that character.

3.2 Neural Approaches

The use of neutral approaches mostly encompasses use of an encoder decoder architecture for the task of sequence labelling. (Faruqui et al., 2015) was one of the first works to explore the use of encoder decoder model for the use of morphological inflection. They use a bidirectional LSTM for the encoder, which takes the character sequence of the surface form as input along with a LSTM decoder. The input to the LSTM decoder is the encoder output along with decoder output from previous time steps. Their model makes no assumptions about morphological processes or rules , and the inputs are simply the individual characters. In one of the experimental settings, they learnt a separate decoder model for each type of inflection (which they call a factored model) and compared it with a model with common decoder for all the inflections. They also used the unlabelled data in order to induce the copying bias in the models in an semi-supervised setting. The highest accuracy was obtained for factored semi-supervised setting on the Wiktionary inflection dataset (Durrett and DeNero, 2013).They found that their models achieve better performances on the previous state of the art models such as plain encoder decoder models without a semi-supervised setting. The shared task of SIGMORPHON - 2016 task is based on morphological re-inflection wherein the root word and

the target MSD tags were given and the task was to transform the root into the target surface form (for which the tags were given). Nine teams provided their submissions for the task. The neural systems outperformed the other systems which were based on linguistic heuristics or classical transduction based models by a large margin. The model by (Kann and Schütze, 2016) which won the competition used bidirectional GRU based encoder decoder architecture with soft attention which was traditionally developed for machine translation. They trained a separate model for each language as opposed to separate model for each type of inflection as was used by (Faruqui et al., 2015). The second ranked system used a separate model for each part of speech type and the model they used was encoder-decoder model with the encoder augmented with a bidirectional LSTM. (Akyürek et al., 2019) design a model called Morse which is a recurrent encoder-decoder model that produces morphological analyses of each word in a sentence. They discovered that forcing the decoder to anticipate both the lemma form and the factored MSDs yields poorer results than decoding simply the MSDs, especially in low-resource settings and for morphologically more complicated languages. They also discovered that when training data is limited and the language includes rich inflections, predicting factored MSDs is preferable to predicting unfactored MSDs, and that utilising related high-resource language data to supplement low-resource language training is beneficial. Their model outputs morphological tags one feature at a time, enabling it to learn unseen/rare tags. At the heart of the system is a unidirectional LSTM. a forward LSTM at the word level to obtain the word embedding for each word; a character level LSTM to get the word embedding for each word read in words to the left of the current word and a backward word-level LSTM to read in words to the right of the current word Take the concatenation of the forward and backward LSTMs as the right-hand side of the current word. additional unidirectional LSTM to encode the expected morphological changes; the context representations tagging of the two words before it.

3.2.1 Transformer based architectures

The second subtask of SIGMORPHON 2019 shared task is on morphological tagging and lemmatization in context. For this subtask, 16 systems were submitted, all of which used neural network models. The most effective systems employ BERT pretrained embeddings for contextual representations and perform versions of multi-headed attention, multi-level encoding, and multiple decoding. (Kondratyuk, 2019) was the best performing system. The authors use the multilingual BERT model and many UDify-developed fine-tuning procedures to achieve remarkable assessment performance on morpho-syntactic tasks. Their findings reveal that fine-tuning multilingual BERT on the concatenation of all of accessible tree banks allows the model to learn cross-lingual information, which improves lemmatization and morphology tagging accuracy when compared to fine-tuning it monolingually. They encode input sentences using the pretrained multilingual BERT cased model, then apply additional word-level and character-level LSTM layers before decoding lemmas and morphological tags simultaneously using simple sequence tagging layers. To identify the edit actions to turn an inflected word to its lemma, lemmatization decoding is regarded as a sequence classification problem, and a feedforward layer is applied to the lemmatizer LSTM final layer to achieve the result. A feed forward layer is used to jointly forecast factored and unfactored MSDs for morphological tagging. Their system is shown in figure 2

(Straka et al., 2019) is also one of the best performing systems for the task. They use modified version of UDipipe 2.0 for the task. After converting the input words to embeddings, three shared bidirectional LSTM layers are performed.Then, softmax classifiers are used to process the output and generate the lemmas and POS tags (morphosyntactic features). The lemmas are created classifying into a series of edit scripts that evaluate the input word form and construct lemmas by executing character-level modifications on the prefix and suffix words. We add pretrained contextual word embeddings (BERT) as another input to the UDIpipe framework. They predict the full POS tag and regularize the model by predicting individual morphological features.

4 Few Shot and Zero Shot Learning

Though the performance of SeqtoSeq models is very impressive in case of high resource scenarios there is a significant drop in the performance when it comes to low resource scenario. Scarcity of labeled examples is one of the core problems in

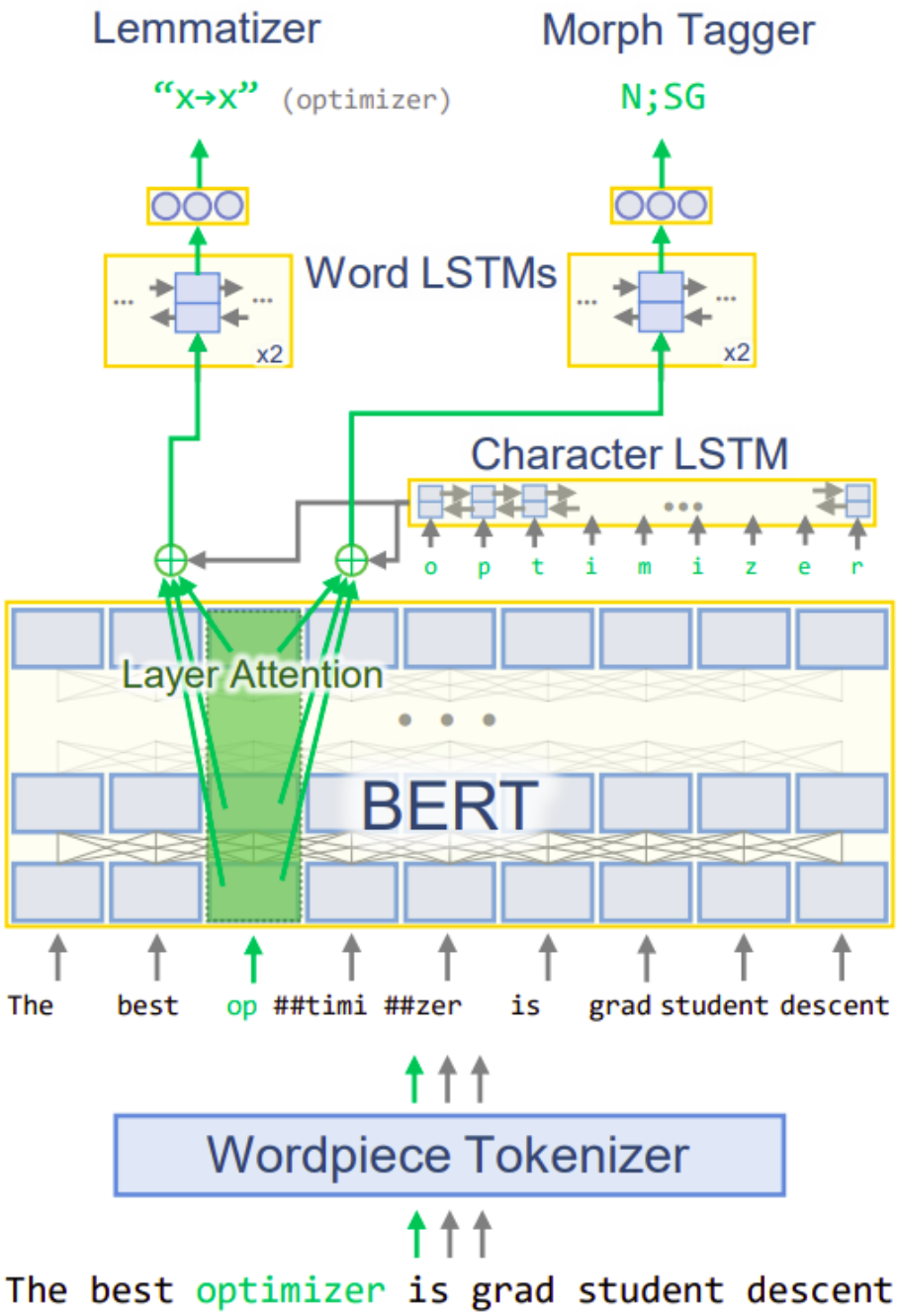

Figure 2: BERT based Morphological analyzer. The mBERT model acts as a multilingual encoder and the word level and character level LSTMs are the decoders

most of the NLP tasks for all languages apart from resource rich languages such as English. This renders the researchers to look into strategies that work well with less amount of data and look beyond multilingual models. One such strategy is to use multilingual models that are trained on a large number of languages and are effectively able to use the linguistic knowledge that is gained from one language (which is probably a high resource language with lots of annotated data) to a low resource language. Currently, the most commonly used technique for few shot transfer in a low resource scenario is the use of continuous cross-lingual representation spaces that encode the linguistic knowledge of multiple languages into a single space. The cross lingual word embeddings

(Ruder et al., 2019) and massively multilingual models that are trained using language modelling objectives (or denoising objectives) have become the mainstay of current few shot transfer techniques (Conneau et al., 2020). (Lauscher et al., 2020) study the effectiveness of zero shot and few shot learning paradigm in the context of low resource languages. They train a model in English and test it on various languages such as Arabic (AR), Basque (EU), (Mandarin) Chinese (ZH), Finnish (FI), Hebrew (HE), Hindi (HI), Italian (IT), Japanese (JA), Korean (KO), Russian (RU), Swedish (SV), and Turkish (TR), etc for 5 tasks: 3 low level tasks (POS tagging, dependency parsing, NER) and 2 high level tasks: Natural language inference (NLI) and QA. They have used multilingual pretrained models mBERT and RoBERTa for these tasks. The authors suggest that presence of few labelled examples of low resource languages can highly enrich the performance of the models in low resource languages especially when the low resource language has lesser similarity to high resource language. The boost in performance is observed mostly in the low level tasks such as NER, POS tagging, etc. The authors also highlight the curse of multilingaulity which occurs when too many languages are used to pretrain the language models and leads to slight dip in the performance. The paper also predicts the performance that a model will have on a low resource language based on the dissimilarity of that language with the source language and number of examples present. Morphology analysis has also been studies in the context of low resource Indian languages in (Saunack et al., 2021) where the main goal is to study effectiveness of cross lingual transfer for the task of lemmatization. The authors use two step attention process in an LSTM based encoder-decoder model. They observe that for most of the Indian Languages, a monolingual model trained on approximately 1000 training samples gives competitive accuracy. They also observe that presence of PoS tags as one of the features benefits the training.

5 South Asian Languages

South Asia is home to a diverse range of languages, including four major linguistic groups and numerous putative linguistic isolates, many of which are severely underserved by contemporary language technology.Furthermore, the languages of South Asia have a lengthy history and have experienced complicated evolution as a result of genetic descent, socio-linguistic interactions, and contact effect. The most important obstacle in developing language technologies, has been resource scatteredness rather than resource scatteredness as argued in (Arora et al., 2022). Most of the languages including endangered ones have a wealth of data to be retrieved from annotated corpora and grammatical descriptions maintained by linguists, if only one is prepared to wrangle and extract unusual and vivid data formats and digitise old texts. The focus of current effort has been on data-scattered languages rather than resource scarce languages. (Kakwani et al., 2020) has been one such effort in the direction of developing resources for South Asian languages. The authors have introduced resources for 11 Indian languages belonging to two major resources. These resources include: (a) large-scale sentence-level monolingual corpora containing a total of 8.8 billion tokens across all 11 languages and Indian English, primarily sourced from news crawls, (b) pre-trained word embeddings based on *FastText*, (c) pre-trained language models based on ALBERT, and (d) multiple NLU evaluation datasets such as: Article Genre Classification, Headline Prediction, Wikipedia Section-Title Prediction, Cloze-style Multiple choice QA, Winograd NLI, The Choice Of Plausible Alternatives (COPA), Named Entity Recognition, Cross-lingual Sentence Retrieval, Paraphrase detection, Discourse Mode Classification, etc.

(McEnery et al., 2000), has also been one of the efforts of past years that was aimed at converting 8-bit language data into Unicode. The corpus is made to support translation and transliteration tools for languages: Bengali, Hindi, Punjabi. (Arora, 2020) is one of the tools available for processing of Indian languages. It contains pre-trained language models: ULMFiT and TransformerXL for 13 Indian languages: Hindi, Bengali, Gujarati, Malayalam, Marathi, Tamil Punjabi, Kannada, Oriya, Sanskrit, Nepali, Urdu. It also conatins support for: Textual Similarity, Data Augmentation,Word Embeddings, Sentence Embeddings, Tokenization and Text Generation in 13 Indic Languages. Workshops such as (Chakravarthi et al., 2021) have been a significant effort in the direction of developing resources for Indian languages. There are resources availabe

for a few languages that are not in usable state currently or amount of labelled data is limited, (Joshi et al., 2020) calls such languages as underdogs. The languages such as Marathi fall in this category. The left behinds category is however a long tailed one and includes several hundreds of languages. With such little resources, bringing them into the digital realm will be a monumental, if not impossible, task. Because there is essentially no unlabeled data to utilise, unsupervised pre-training approaches just make the 'poor' poorer. The scarping-bys are the ones for which with a significant amount of unlabeled data, they may be in a stronger position in the 'race' in a couple of years. This endeavour, however, will need a well-coordinated campaign that raises awareness of these languages while simultaneously spurring a concerted effort to gather tagged datasets for them, which they now lack. Indian languages that fall in the category are:Malayalam, Bhojpuri, Nepali, Doteli, Gujarati, Newar, Dzongkha, Maithili, Tulu, Kannada, Odia, Kashmiri, Romani, Pashto, Bishnupriya Manipuri, Divehi, Sindhi, Tibetan, Pali, Sinhala, Santali, Assamese, Telugu. Hopefuls are the languages for which a modest amount of labelled datasets have been gathered, indicating that scholars and language support networks are working to keep them alive in the digital world. In a few years, promising NLP solutions for these languages might be developed, Konkani, Sanskrit, Punjabi fall in this category. In the case of current South Asian languages, there has been recent diversity as a result of collaborative initiatives, such as an impending shared task on dependency parsing at the WILDRE 2022 workshop based on new treebanks.

6 Approaches for low resource settings

Languages are categorised according to their structural and semantic characteristics in the discipline of linguistic typology. A database of typological traits between languages has been developed as a result of extensive work such as (Dryer and Haspelmath, 2013). Given that there are so few categories of comparable scope, such documentation becomes crucial. There has been research in the field of NLP showing the value of using typological information to direct model creation (Ponti et al., 2019). Additionally, it has been demonstrated that transfer learning of resource-rich languages to resource-poor languages performs better if the two languages share similar typological traits as shown in (Pires et al., 2019).The authors demonstrate that Multilingual BERT (M-BERT) is surprisingly effective at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. M-BERT is a single language model that was pre-trained from monolingual corpora in 104 languages. They demonstrate through a large number of probing experiments that transfer is possible even to languages with different scripts, that transfer is most effective when languages are typologically similar, that monolingual corpora can be used to train code-switching models, and that the model can identify translation pairs. Because M-BERT allows for a very simple method of zero-shot cross-lingual model transfer, it is especially well suited to this probing study. We refine the model using task-specific supervised training data from one language, and then evaluate that task in a different language, allowing us to observe the ways in which the model generalises information across languages. M-BERT performs cross-lingual generalisation surprisingly effectively, according to results. More crucially, the study offers the findings of a series of probing experiments meant to explore alternative ideas about how the model may execute this transfer.

Even while all languages have profound, abstract aspects in common, the patterns used in everyday, surface-level writing might differ greatly. The development of strong, multilingually applicable NLP technology has been hampered by this cross-lingual diversity, and as a result, current NLP is still primarily restricted to a small number of resource-rich languages. Most existing algorithms are far from being language-agnostic in their architecture design, training, and hyperparameter tuning, and frequently unintentionally introduce language-specific biases (such as models being adapted to handle morphologically poor languages as highlighted in Bender (2011)). A condition that cannot be satisfied for the majority of the world's languages is that the bulk of modern machine learning models rely on supervision from (huge quantities of) labelled data. Cross-lingual variance is restricted and far from random, according to analysis of cross-linguistic patterns as highlighted in (Greenberg, 1963). It is true that typological traits can be interdependent: The existence

of one feature may suggest the existence of another (in one direction or both). As contrast to unrestricted universals, which define features shared universally by all languages, this dependency is referred to as constrained universal. Such restricted or unrestricted typological universals are seldom absolute (that is, exceptionless), but rather trends, which is why they are referred to as "statistical" (Corbett 2010).

Languages with abundant resources can transmit linguistic information to languages with limited resources; these are referred to as source languages and target languages, respectively. It is difficult to transfer languages because we must match word sequences with various lexica and word ordering, as well as syntactic trees with various (an-isomorphic) structures (Ponti et al., 2018). Because of this, it is usually necessary to change the information from the source languages to fit the characteristics of the destination languages. Annotation projection, (de)lexicalized model transfer, and translation are techniques developed for language transfer. From the data in many languages, NLP models may be collaboratively trained. This method generally outperforms language-specific monolingual models since it can make use of more (although coarser) data in addition to helping applications that are inherently multilingual, such as Neural Machine Translation and Information Extraction (Ammar et al., 2016).This is especially true in situations when a target language or all target languages are resource-lean, such as in code-switching settings (Khapra et al., 2011). Sharing parameters is a crucial tactic for multilingual cooperative learning (Johnson et al., 2017). More specifically, input and hidden representations in cutting-edge neural networks can either be private (language-specific) or shared across languages. The linking of a network component's parameters across languages, such as word embeddings, character embeddings, hidden layers, or the attention mechanism leads to shared representations. Providing details about the language of the current text in the form of input language vectors is another widely used technique in multilingual joint modelling. This should make it easier to adapt the joint model to other languages, goes the reasoning. In neural language modelling tasks, or NMT tasks, these vectors can be learnt from beginning to end.

Similar words—regardless of language—are represented similarly in multilingual word embeddings. Many techniques have been developed to produce multilingual word embeddings. Monolingual mapping generates independent monolingual representations and thereby learns a linear map between a source language and a target language with the help of a bilingual lexicon (Mikolov, Le, and Sutskever 2013) or in an unsupervised fashion through adversarial networks (Conneau et al., 2017). Words from other languages are combined with their contexts in pseudo-cross-lingual techniques, which then use this mixed corpus to create representations. Wiktionary (Xiao and Guo, 2014) or machine translation are used as replacements.

7 SIGMORPHON Shared Tasks

SIGMORPHON (Pimentel et al., 2021) has been one of the venues that has been arranging workshops,shared tasks and competitions on Morphology, Phonology in recent times. The shared tasks provide researchers morphological data to apply and test deep learning models for morphology related tasks. The SIGMORPHON shared task 2019 (McCarthy et al., 2019) is examination of context and cross-lingual transfer The study of transfer learning in morphology was conducted. as well as inflections between 100 different language pairings in 66 languages as morphosyntactic description and contextual lemmatization. It consisted of two challenges, first was transfer of morphological inflection knowledge form high resource language to other. The second challenge was on lemmatization as well as morphological feature analysis in context ie. lemmatize each word in an unannotated sentence and tag them with a morphosyntactic description. For the first task, the data for all languages except four (Basque, Kurmanji, Murrinhpatha, and Sorani) comes from English Wiktionary, a huge multilingual crowd-sourced dictionary including morphological paradigms for numerous lemmata. The task received 30 submissions—14 for challenge 1 and 16 for challenge 2— from 23 teams. The University of Alberta (UAlberta) conducted a targeted examination on four language pairings, using external cognate lists to train cognate-projection systems.Two approaches were considered: one that trained a high-resource neural encoder-decoder and projected test data into the HRL, and the other that projected HRL data into the LRL and trained a combined system. AX-Semantics utilised low-

and high-resource data to train a seq2seq encoder-decoder model, using domain adaptation strategies to focus later epochs on the target language as an alternative. Cross-lingual training yielded moderate increases in task 1, with gains positively correlated with the linguistic closeness of the two languages. For task 2, by employing multi-lingual BERT embeddings fine-tuned on a concatenation of all accessible languages, Charles-Saarland was able to obtain the highest overall tagging accuracy, successfully carrying the cross-lingual aim of Task 1 into Task 2.

SIGMORPHON Shared Task 0, 2020 (Vylomova et al., 2020) was on Typologically Diverse Morphological Inflection. Data from 45 languages and just five language families were used to train the systems, which were then fine-tuned with data from another 45 languages and ten language families (for a total of 13 languages) before being tested on all 90 languages. The task received 22 systems (19 neural) from ten teams, with all four winner systems being neural. For low-resource languages, the majority of teams focused on the value of data hallucination and augmentation, ensembles, and multilingual training. The task considered three dimensions of morphological variation: fusion, inflectional synthesis, and position of case affixes. Fusion refers to the degree to which morphemes bind to one another and languages can vary from strictly isolating to concatenative. The most frequent system is concatenative morphology, which may be found all around the world. Inflectional synthesis refers to whether grammatical categories like gender, person, number,tense, aspect, modality voice or agreement are expressed as affixes (synthetic) or individual words (analytic) as markers. Affixes can variably occur as prefixes, suffixes, infixes or post-positions. Neural baselines were based on a neural transducer (Wu and Cotterell, 2019), which was essentially a hard monotonic attention model and vanilla transformer model adapted for character level language modelling. There were around 4 winning systems. One of the submissions in the task also manually designed finite state grammars for 4 languages and found them to have superior accuracy but noted that the accuracy came on the top of significant person hours. Some submissions also demonstrated utility of data hallucination.

The shared task 2 of SIGMORPHON 2020 (Kann et al., 2020) was unsupervised morphological paradigm completion. The task was to design a system that takes raw text and a list of lemmas as input, and output all inflected forms, also known as, morphological paradigm, of each lemma. There are various sub-tasks to this task: To begin, a system must determine which words in the corpus are part of the same paradigm, the second step is to determine the paradigm's form, this necessitates determining which forms of various lemmas convey the same morphosyntactic properties, despite the fact that they are not produced in the same way, third the system also needs to produce all the paradigms not mentioned in the corpus. The baseline used the following steps: edit tree retrieval, additional lemma retrieval, paradigm size discovery, and inflection generation. The shared task 0 of SIGMORPHON 2021 (Pimentel et al., 2021) focused on typo-logical diversity and cross-lingual variation of morphosyntactic features. Transformer-based models outperformed traditional models in the majority of languages, reaching higher than 90% accuracy in more than half of the languages. They observed that, the majority of system errors are caused by allomorphy, honorificity, and form variation. The systems modelled morphological inflection as series of inserting, deleting, and/or replacing fixed characters (in no specific order). The shared task 2 (Wiemerslage et al., 2021), focused on Unsupervised Morphological Paradigm Clustering. The authors made corpora available for five development and nine test languages, as well as gold partial paradigms for testing. A supervised lemmatizer outperformed all of the systems, indicating that there is still opportunity for improvement. According to the authors, a good unsupervised paradigm clustering method takes use of common characteristics in a language's inflectional morphology while neglecting regular contextual and derivational patterns. There were two types of system submission to the task, similarity-based systems which used various combinations of orthographic and embedding-based similarity metrics for word forms, as well as clustering algorithms such as k-means and agglomerative clustering, Methods based on grammar extract grammars or rules from the data and use them directly to clustering, or partition words into stems and affixes before grouping forms that share a stem (lemma) into paradigms. The authors also note that all methods produce the greatest results for

English, Spanish, and Bulgarian, in that order. These three languages are all heavily suffixing, although inflection is usually expressed with just one morpheme.

8 Conclusion

This survey provides a review of various machine learning approaches (supervised and unsupervised) that are used for the task of morphology analysis. It provides a survey of unsupervised and geometric approaches such as Minimum description length that are used for the task of discovery of morphological structures. We also discussed various neural as well as transformer approaches that have been used for the task of morphology analysis. We then dived into approaches that are used in low resource settings. We finally provided a survey of the current state of resources for South Asian languages along with an overview of SIGMORPHON shared tasks.

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.

Ekin Akyürek, Erenay Dayanık, and Deniz Yuret. 2019. Morphological analysis using a sequence decoder. *Transactions of the Association for Computational Linguistics*, 7:567–579.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

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.

Gaurav Arora. 2020. iNLTK: Natural language toolkit for indic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 66–71, Online. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

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.

Erwin Chan. 2006. Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 69–78, New York City, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. pages 8440–8451.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Mathias Creutz and Krista Lagus. 2004. Unsupervised discovery of morphemes.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2015. Morphological inflection generation using character sequence to sequence learning. *CoRR*, abs/1512.06110.

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12:353–371.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2005. Interpolating between types and tokens by estimating power-law generators. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 459–466, Cambridge, MA, USA. MIT Press.

Joseph Harold Greenberg. 1963. Universals of language.

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.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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, Online. Association for Computational Linguistics.

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 51–62, Online. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Mitesh M Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011. Together we can: Bilingual bootstrapping for wsd. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 561–569.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. 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.

Ling Liu. 2021. Computational morphology with neural network approaches. *arXiv preprint arXiv:2105.09404*.

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.

Anthony McEnery, Paul Baker, Rob Gaizauskas, and Hamish Cunningham. 2000. EMILLE: building a corpus of South Asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*, University of Exeter, UK.

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.

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.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Ponti, Roi Reichart, Anna-Leena Korhonen, and Ivan Vulic. 2018. Isomorphic transfer of syntactic structures in cross-lingual nlp. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Kumar Saunack, Kumar Saurav, and Pushpak Bhattacharyya. 2021. How low is too low? a monolingual take on lemmatisation in Indian languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4088–4094, Online. Association for Computational Linguistics.

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.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.

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.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. 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, Online. Association for Computational Linguistics.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

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.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.