International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Part of Speech and Morph Category Prediction for Gujarati

Jatayu Baxi [1*], Om Soni [2], Brijesh Bhatt [3]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: This paper presents a novel approach for the prediction of Part of Speech (POS) category and Morphological features for the Gujarati language. POS tagging and Morphological analysis are foundation level tasks in almost all Natural Language Processing (NLP) applications. For the low resource and morphologically rich languages like Gujarati, the task becomes more challenging. In this work, we explore transformer based pre-trained models for the underlying task. We propose 4 different models for the prediction of POS category and Morph features. Along with the prediction of POS tagging and Morphological features individually, this work also explores the linguistic relationship between these features and proposes a single joint model for the prediction of POS-MORPH features. The joint model achieves F1 score of 0.98 and outperforms individual models.

Keywords: *NLP, Transformer, BERT, Gujarati, Deep Learning*

1. Introduction

POS tagging and morphological analysis are fundamental tasks in NLP. They help in understanding the grammatical structure and semantic meaning of a word in a better way. POS tagging is a linguistic analysis task in natural language processing where words in a sentence are assigned specific grammatical categories such as nouns, verbs, adjectives etc to understand their syntactic and semantic roles in a sentence[1]. On the other hand, a morph analyser separates root and suffix part and assigning grammatical features to the inflected word[2]. Table 1 shows the example of POS tagging and morphological analysis in English and Gujarati languages. For example, in the sentence `Alex goes for a walk.', consider word goes. The POS tag for the word `goes' is Verb and the morphological tag is: Go (root form), present tense, third person singular.

In this paper, we present POS tagger and Morphological analyzer for the Gujarati language. Developing efficient POS tagging and morphological analyzer is a challenging task due to language specific complexities in the word formation process. One major issue is ambiguity and context sensitivity. Multiple meanings of the same word based on the context creates problems in accurate POS tagging and morphological analysis. For example, consider word **run** in the sentences `I like to **run** in the park every morning' and `The engine is designed to **run** smoothly.' In the first sentence the word **run** is used as verb and it means the act of running while in the second sentence it is used as a noun and it means continuous movement The same word may have different POS or morphological characteristics in different contexts. Another issue is related to the annotated data. For efficient POS tagging and morphological analysis systems, it is important to have large annotated training corpus. For many languages, especially low-resource languages, the availability of such annotated data is limited. Due to the variations in the linguistic structure, morphology and writing style of different languages, the adaption of common POS tagging and morphological analyzer system is difficult.

[1,2,3] *Department of Computer Engineering, Dharmsinh Desai University, Nadiad (Gujarat)*
1 ORCID ID: 0000-0001-5377-7161
2 ORCID ID: 0009-0003-3164-2364
3 ORCID ID: 0000-0002-7934-7992
**Corresponding author Email ID : jatayubaxi.ce@ddu.ac.in*

[1] https://en.wikipedia.org/wiki/Part-of-speech_tagging
[2] https://en.wikipedia.org/wiki/Morphology_linguistics

Sentence	POS tagging	Morphological Analysis
The sun sets beautifully.	The/DET sun/NN sets/VB beautifully/ADV	The/[The] sun/[sun, singular] sets/[set, present _ tense] beautifully/[beautiful].
કિશોરે જવાબ આપ્યો.	કિશોરે/N_NNP જવાબ/N_NN આપ્યો/V _ VM .	કિશોરે/ [કિશોર,N_ NNP, Male] જવાબ/[જવાબ, NN, Singular, Nominative, Male] આપ્યો/[આપવું, VB, Past _ Tense]

Table 1: POS tagging and Morphological category prediction examples in English and Gujarati

Gujarati is one of the major languages in India. It is derived from ancient language Sanskrit. Gujarati has 62 million speakers worldwide[3]. Gujarati is a rich language consisting of 34 consonants and 13 vowels. Gujarati follows a subject-object-verb (SOV) word order. Gujarati is morphologically rich language. It has number of inflections and derivations which adds to the complexity of the morphology. In contrast to the other Indian languages such as Hindi, Gujarati has 3 genders; masculine, feminine and neuter. Noun phrases in Gujarati often display gender and case inflections, while verb conjugations vary based on tense, aspect, and person. The highly inflectional nature of Gujarati presents unique challenges in developing POS tagger and morphological analyzer tool. Below are some of the language specific morphology challenges for the Gujarati language [1] :

- Agglutinative Nature and word compounding: In Gujarati, multiple morphemes are often combined to form a single word. This makes it challenging to separate and analyze individual morphemes accurately. For example, word **વૃદ્ધાવસ્થા** (Old age) is made up from two morphemes **વૃદ્ધ + અવસ્થા**. Due to this word compounding, it is challenging to break down and tag individual components.
- Lack of Clear Word Boundaries: Due to this, it becomes difficult to determine where one word ends and another begins. For example, consider words **ફરિયાદ** and **ફરી યાદ**.
- Polysynthetic Feature: Due to polysynthetic features, a single word can convey a complex meaning through the combination of multiple affixes and morphemes. Consider a word **છોકરાઓનું** which is made up from a morpheme **છોકરા**, plural marker **ઓ** and case marker **નું**.
- Limited Resources: Compared to other major Indian languages, Gujarati has limited linguistic resources, such as annotated corpora and morphological dictionaries.

For Gujarati language, some research work has been done in this area. The existing systems focus on predicting POS and Morph features in isolation. We analyze that there is strong relationship between POS category of a word and its morphological features in Gujarati. Instead of predicting POS category and morphological features in isolation, we can combine them together and achieve better performance. Below are some observations on how POS-Morph features help each other:

- Context Enrichment: Providing the POS category offers valuable contextual information about the word's role within a sentence. It helps the model understand syntactic relationships, influencing the word's potential morphological properties. Also, incorporating POS information enriches the word embeddings by providing additional context.
- POS-Morphology Correlation: Often, specific POS categories correlate with certain morphological features. For instance, verbs may exhibit distinct morphological patterns compared to nouns or adjectives. Having the POS category as an input guides the model to focus on relevant morphological features associated with that specific POS tag, potentially improving the accuracy of predictions.
- Resolving Ambiguity: When a word might have multiple morphological interpretations. POS information acts as a disambiguating factor, narrowing down the possible morphological features.

Consider two Gujarati sentences in Table 2. The example above illustrates that the word `**નાની**' has two meanings: When it is acting as noun, it means `**grandmother**' and

[3] https://en.wikipedia.org/wiki/Gujarati_language

when it acts as an adjective, it means `**small**'. If we use a model which only predicts POS or morph feature then it may not give correct output due to this ambiguity. If we use a model which jointly predicts POS and morph feature then it will generate proper output. It means that if we know that the word belongs to Noun category then the model will accurately predict morph features as gender, number and case marker and if the word belongs to adjective category, then it will predict features gender, number and type. Similarly, knowing the morph features gender, number and case marker, the model can accurately predict POS category as Noun. For instance, the word `**નાની**' has three variations `**નાની**', `**નાનો**' and `**નાનું**' when it is used as adjective depending upon the gender of the noun it follows. but it has only one form `**નાની**' when used as noun.

Gujarati Sentence	**Transliteration**	**English Translation**
મેં નાની બચત યોજનામાં રોકાણ કર્યું છે.	Mēṁ nānī bacata yōjanāmāṁ rōkāṇa karyuṁ chē.	I have invested in small savings plan.
મારા નાની ગામડે રહે છે	Mārā nānī gāmaḍē rahē chē	My Grandmother lives in a village.

Table 2 : Example of POS-Morph ambiguity in Gujarati

Below are major research contributions of this paper:

- We analyze morph categories, POS tags and interdependence between them in the Gujarati language.
- We propose a model which jointly predicts morphological category and POS tag of a word in the sentence for the Gujarati language.

The remaining of the paper is organized as follows. Section 2 describes the related work. In section 3, we provide details about the dataset. In Section 4, the details about the proposed models are explored. In section 5, experiments and results are discussed followed by the conclusion.

2. Related Work

In this section, we highlight the survey of the existing work in the field of POS tagging and morphological analysis. After discussing general survey, we also discuss the work done specifically for the Gujarati language and discuss research gaps.

The initial efforts for the development of morphological analyzer were based on stemmer and finite state transducer approaches. In the stemmer-based approach, stemming rules are used to obtain root word and identification of grammatical features [2]. Later on, researchers used finite state transducer to encode these rules using the concepts like two level morphology [3]. In the decade of 2000-2010, paradigm-based approaches and unsupervised approaches were explored [4, 5]. These approaches had a limitation of manual rule building and they often failed to provide accurate results due to language ambiguities. After 2010, various machine learning and statistical approaches were investigated for this problem. The machine learning approaches performed better than the traditional rule-based approaches but they required heavy feature engineering [6, 7]. After the introduction of deep learning-based models, the feature engineering was not required. Due to this advantage, deep learning-based morph analyzers for different languages have been created [8, 9, 10]. In the recent times, transformer-based approaches have become popular for almost all NLP tasks. The core of transformer-based models such as BERT are multilingual pre-trained models. Such models are trained on the large corpus and then they can be fine-tuned for the specific task. For the task of morphological analysis and lemmatization, transformer-based approaches have been experimented [10, 11, 12].

The efforts of developing POS tagger dates back to 1992 [13]. Initially, the rule-based taggers emerged and that heavily depended on handcrafted linguistic rules. Though the rule-based approaches can be accurate for specific languages or domains, they can become complex and challenging to maintain as languages evolve or when handling ambiguous cases [14, 15]. Subsequently, statistical models such as Hidden Markov Model (HMM) gained popularity. They work by estimating probabilities for predicting POS category. This approach often yields superior performance in POS tagging tasks compared to traditional models due to its ability to capture complex linguistic patterns and contextual dependencies [16, 17]. With the advancements in the machine learning techniques, various data driven approaches which utilized techniques like decision tree, naive bayes etc. were explored. Later on, researchers used various neural network and deep learning-based architectures for this task. Deep learning models have numerous advantages such as feature learning, contextual understanding, generalization etc. However, the deep learning models require substantial training data and computational resources for training and inference for efficient results. Some noteworthy contributions of developing POS tagger using deep learning models are [18, 19, 20, 21]. In the recent years, for the POS tagging, the state-of-the-art results are obtained by using power of pre-trained transformer-based models [22, 23] .

For the Gujarati language, some research works have been carried in the field of POS tagging. In [24], authors have compared various approaches for Gujarati POS tagging.

The work [25] discusses CRF based POS tagger. In [26], authors have proposed hybrid method for the POS tagging using LSTM and linguistic rules. The work in [27] describes LSTM based POS tagger for the Gujarati language.

The development of morphological analyzer for Gujarati is a challenging task. Compared to other languages, less work is reported in Gujarati. In [28], authors have developed rule based morphological analyzer for Gujarati by hand crafting of suffix rules. A Bi-LSTM based model for Gujarati morphological analysis is proposed in [29]. The model was improved in the work [30] by selecting a different label representation approach. These works use Gujmorph dataset [31].

The rule-based POS taggers and morph analyzers rely heavily on the hand-crafted suffix rules and linguistic resources like lemma dictionaries. Due to the ambiguity in the word formation rules in Gujarati, such approaches do not produce promising results. Also, it is difficult to create or maintain such rules as it requires lot of human efforts. The standard machine learning based approaches require manual feature engineering which is complex task for the highly inflectional languages. Existing works typically treat POS tagging and morphological analysis as distinct tasks. Deep learning models like Bi-LSTM are used in [30] and [27] but more advanced transformer based pre-trained models are not explored till date. Our work aims to address these shortcomings by proposing BERT-based pre-trained models. This novel architecture performs joint prediction of POS and morphological features for the Gujarati language. Leveraging linguistic knowledge from pre-trained models, our approach eliminates the need for manual feature engineering or rule crafting, providing a more efficient solution.

3. Dataset

In this section, we describe the dataset that we have used to train our models. We describe how two different datasets are combined to create a dataset for the joint prediction of POS-Morph features.

Name	Target Language(s)
Unimoprh [34]	169 languages
UD-Treebank [35]	148 languages
Mighty-Morph [36]	English,Germen, Hebrew, Turkish
Neural-Morphology-Dataset [37]	Germen, Finnish and 20 others
MorphyNet [38]	Russian, Hungarian and 13 others

Table 3: Details of various morphological datasets

Most of the work done for POS tagging makes use of UD treebank [32]. The morphological analyzer models use Unimorph [33] dataset. In Table 3, we have listed other popular data sets for POs tagging and morphological analysis in various languages. In our proposed approach, we use transformer based pre-trained models. These models are already pre-trained on large corpus. We need to fine-tune them on our task specific dataset. Acquisition of POS-Morph annotated dataset for Gujarati is challenging task. For the task of joint prediction of POS-morph features, we require annotated data in such a format where each word is annotated with POS category as well as Morph features. For Gujarati language, we have two separate datasets; Gujarati ILCI-II POS tagged dataset developed by TDIL[4] and Gujarati dataset in the Unimorph schema [34]. The POS tags are defined as per Bureau of Indian Standards (BIS) tag set. For our proposed work, we combined two datasets and created POS-Morph annotated dataset such that for a given word, both the POS tag and morph tags are annotated together. Table 4 shows the details about Gujarati POS-Morph dataset.

[4] http://www.tdil-dc.in

Number of Sentences	30000
Number of unique POS tagged words	50183
Number of unique Morph	50183
Format of the dataset	Word/POS_ Tag/Morph_feature 1;Morph_ feature2; . . ; Morph_ Feature_ n
Example Sentence	સૂકાં પાંદડા અને ધૂળ જમીન પરથી ઉઠી.
Example tagged sentence from dataset	સૂકાં/JJ/સૂકું/ADJ;PL;LGSPEC02 પાંદડાં/N_NN/પાંદડું/N;NOM;NEUT;PL અને/CC_CCD\NA\NA થોડી/QT_ QTF/NA/NA ધૂળ/N_ NN/ ધૂળ/N;NOM;FEM;SG જમીન/N_ NN/ જમીન/N;NOM;FEM;SG પરથી/PSP/NA/NA ઊઠી/V_ VM/NA/NA

Table 4 : Details about Gujarati POS-Morph Dataset

4. Proposed Models

This section describes architectures of our proposed models. We first discuss the architectures of standalone POS and morph models and then present the joint POS-Morph model.

As discussed in the previous section, most of the present models for POs tagging and morphological analysis are based on either traditional approaches or deep neural network architectures such as RNN, LSTM. In the current times, utilization of pre-trained models for various NLP tasks has emerged as very good strategy. The task of POS tagging and morphological analysis depend on understanding linguistic features of a language. Training a model from scratch demands substantial labeled datasets and computational resources. Pre-trained models, on the other hand, are already trained on extensive corpora and understand linguistic patterns. In our experiments, we use transformer based pre-trained models. These models capture contextual information and semantic relationships. We can use pre-trained models as it is and apply fine-tuning which involves adapting these models to specific tasks or domains using smaller, task-specific datasets. Since our work is for Gujarati language, we require pre-trained model which supports multiple languages. One such model is multilingual BERT(m-BERT) [39]. has capacity to comprehend and represent text in multiple languages. It is developed by Google AI, mBERT is an extension of the original BERT model, trained on a diverse corpus over a hundred languages. m-BERT leverages a single model with a unified vocabulary and parameters, allowing it to encode and interpret text in numerous languages simultaneously. For Indian languages, IndicBERT model has been developed [40]. It is a multilingual ALBERT model pre-trained exclusively on 12 major Indian languages. IndicBERT has much fewer parameters than other multilingual models (mBERT, XLM-R etc.) while it also achieves a performance on-par or better than these models. The 12 languages covered by IndicBERT are: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu. For the Gujarati language, Gujarati-BERT model has been developed by [41]. GujaratiBERT is a Gujarati BERT model trained on publicly available Gujarati monolingual datasets. This model is improvement over the traditional indic-BERT model. The model is evaluated on various tasks such as sentiment analysis, named entity recognition etc. In our experiments, we use Gujarati BERT model.

We propose four distinct models for predicting Gujarati POS and Morphological features. The first two models focus on individual predictions for POS category and morphological features respectively. However, our observations in Section 1 indicate a mutual influence between POS and Morph features. To explore this relationship, the third model takes POS as input and predicts morphological features, to assess the impact of POS category on morphological predictions. In the fourth and final model, we jointly predict both POS and Morph features. The following subsections elaborate on the detailed architecture of each model.

A. Individual prediction of POS category and morphological features.

B. Individual prediction of POS category and morphological features.

C. Prediction of morphological features with POS input to measure its influence.

D. Joint prediction of POS and Morph features.

4.1. Model A and B : Standalone models for POS and Morph prediction

Figure 1 shows the model architecture for the standalone models for predicting POS category and Morph category. This model is a custom token classification model built on

top of BERT, specifically using BertForTokenClassification. Below are common components of the model :

- Embeddings : Represents words as dense vectors of size 768.
- Position embeddings: Represents the position of each token in the sequence.
- Token type embeddings: Represents the segment or type of token.
- BERT Encoder : Comprises multiple layers (12 in this case) of BertLayer.
- Liner Layers : Linear layer with input size 768 and output size n. This layer is used for token classification, taking the contextualized token representations from BERT and predicting a label for each token. The output dimension n indicates it predicts among n different classes or labels for each token. The labels are POS tags in case of Model A and morph categories in case of model B.
- Softmax Layer : Softmax layer is used to convert output scores into probabilities.

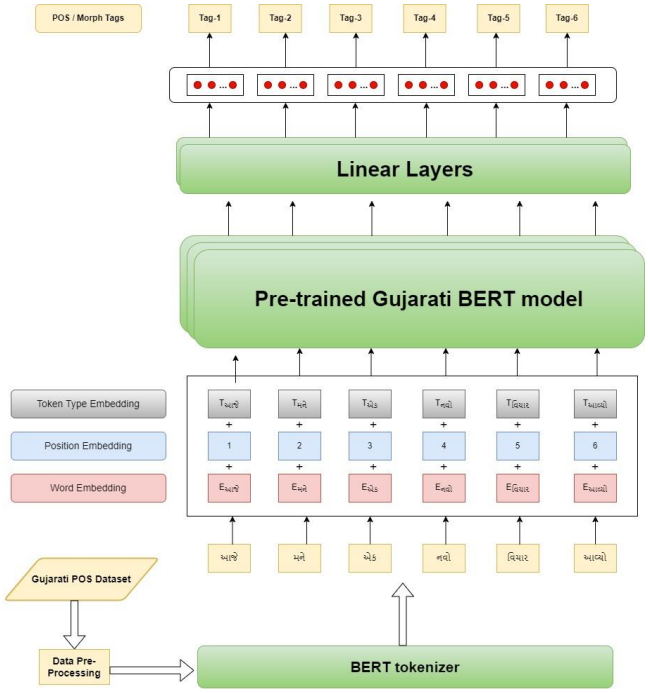

Fig 1: Architecture of Model A and B: Standalone model for the prediction of POS category or Morph category

4.2. Model C : Morph prediction using POS support

Figure 2(a) shows the architecture of the model C, which predicts morph features with POS support. In this model, we keep POS information available in the dataset along with the morphological tags. So, while predicting morphological tag, the model has information about POS category of the word. The model works in the similar way as model A and B with the slight change. Since the dataset contains POS information, the output of the pre-trained BERT has information about both POS and morph tags for a word. This information is passed to the linear layers. This information is propagated to multiple nodes responsible for predicting different morphological features.

4.3. Model D: Joint models for the prediction of POS-Morph features

In this model, we predict POS and morph features together in a joint manner. Figure 2(b) shows the architecture of the proposed joint model. The model is similar to previous models, except the final linear classification layer which has now 67 nodes. 36 nodes for POS tagging and remaining nodes for morphological tags. The custom output layer comprises nodes tailored for specific linguistic features, including 1 node each for POS tagging, Gender, Number, Type, Person, Case, Tense, Aspect, and others. Each node within the output layer is designed to predict and assign values for the corresponding linguistic feature. For instance, the POS node predicts values for Part-of-Speech such as (Noun, Adjective, etc.) and the Gender node predicts values such as MASC, FAM, NEUT, and so forth. During the training phase, a particular emphasis is placed on refining the custom output layer to enhance its predictive capabilities for POS tagging and Morphological analysis.

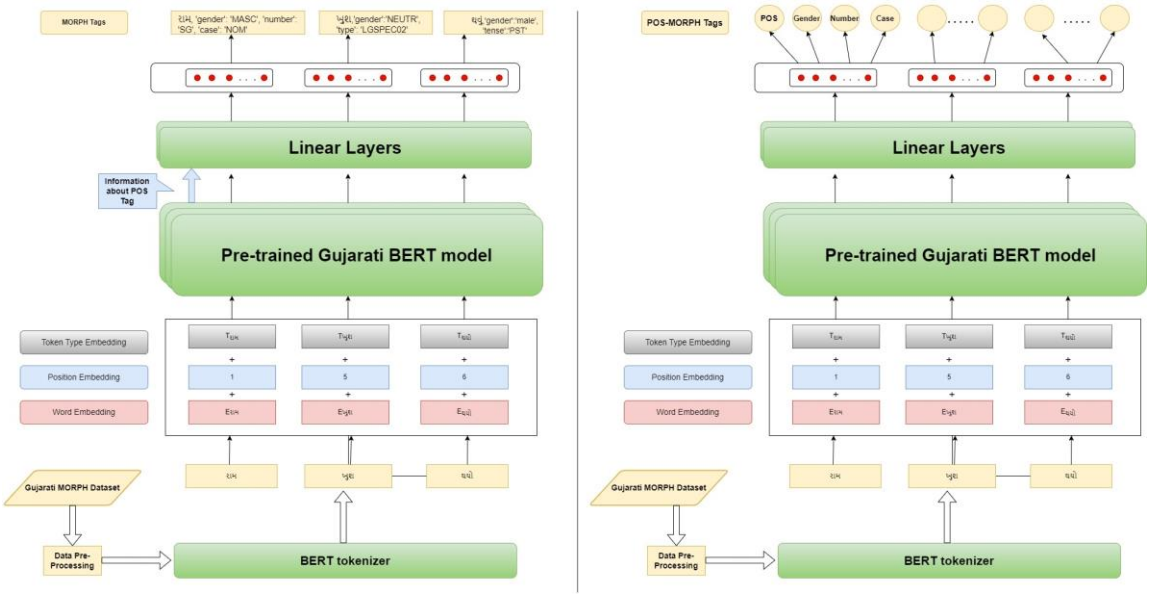

Fig 2 : (a) Architecture of Model C: Morph prediction with POS support **(b)** Architecture of Model D: Joint POS-Morph model

5. Experiments and Results

This section highlights experiment setup and configurations of the hyper-parameters. We also discuss in detail how the joint POS-Morph model performs better than standalone models through analysis of the training. we also review the results obtained. We also explore some good and bad examples to understand the behaviour of the proposed model in a better way.

5.1. Experiment Setup

In our experiments, we use Gujarati BERT pre-trained model and fine-tune it on our labelled dataset described in section 3. The overall experiment setup remains the same for all 4 experiments. Below are the steps for the fine-tuning process.

- Load a pre-trained Gujarati BERT model for token classification.
- Tokenize the custom dataset using the Gujarati BERT tokenizer. Convert tokens and labels tags into numerical formats that can be fed into the model.
- Set up data loaders for training and validation sets. For our experiment, we use 70:30 (23996 sentences for training and 3999 sentences for testing) ratio for training and testing.
- Define a customised output layer of linear nodes to employ the Gujarati BERT model for the prediction of Part-of-Speech (POS) tagging and Morphological analysis.
- Define hyper-parameters and fine-tune the model.

We use the batch size of 8, 15 epochs and learning rate of 2e - 5 as hyper-parameters. below hyper parameters. We carry out our experiments on T4 GPU provided by Google Colab. We use precision, recall and F1-score measures as evaluation matrices. We have evaluated results on the below validation matrices:

- Individual Feature Metrics for Each Word: For each feature (e.g., POS, Gender, Number, Type, Person, Case, Tense, Aspect, etc.) and for each word calculate precision, recall, and accuracy by comparing predicted values to actual values.
- Morphological Overall Metrics: Combine the results from all the individual features to calculate overall accuracy, precision, and recall for morphological features by considering total true and false prediction from all morph features.
- POS wise measures: Given the POS category, compute validation metrics, i.e., specifically for words categorized as `Noun', compute metrics for all POS and Morph features; similarly, for words categorised as `Adjective' and other POS categories, calculate respective metrics.

5.2. Results

Table 5 shows the results for all the proposed models. Table 6 shows the breakdown of the joint model results morph category wise. From the result tables, we observe that the best results for POS tagging and morph category prediction are F1-score of 0.96 and 0.98 respectively. These results are achieved in a model which jointly predicts POS and morph features. Figure 3 and 4 shows the sample loss vs epoch graph for standalone POS tagging model and for the joint model. It is seen from the graph that the

models are converged well which indicates proper training.

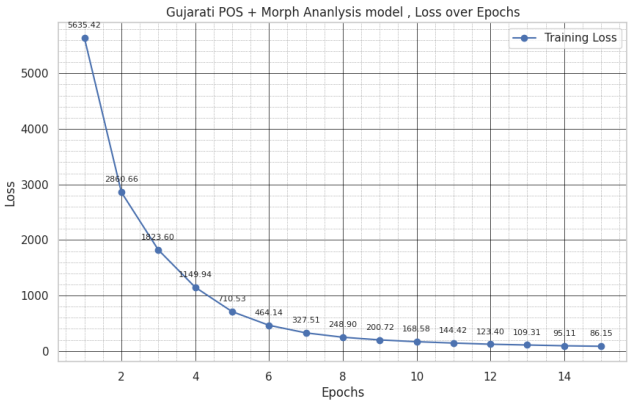

Fig 3: Loss vs Epoch graph for only POS tagging model

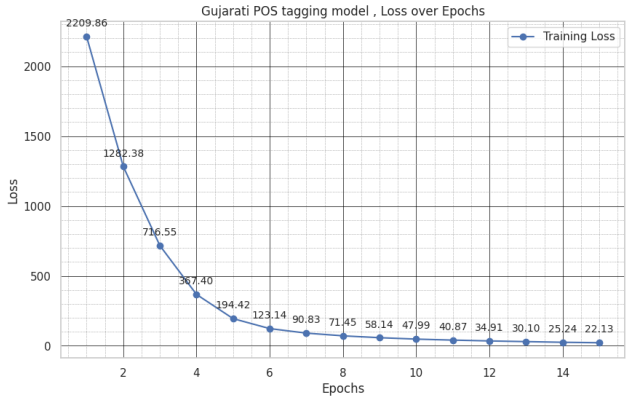

Fig 4: Loss vs Epoch graph for joint POS-Morph model

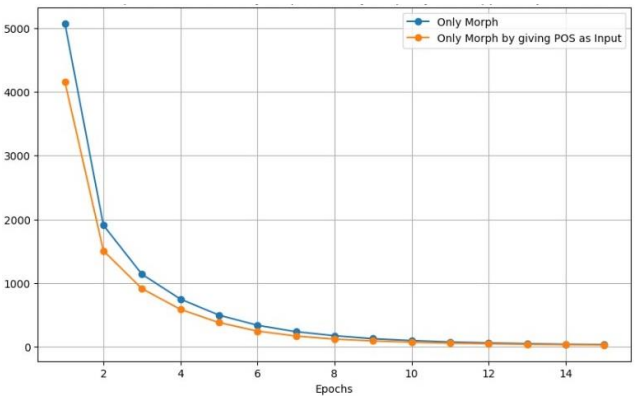

Fig 5: Loss graph for morph analysis model with POS input

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Only POS	0.95	0.95	0.95	0.95
Only Morph	0.97	0.97	0.97	0.97
Morph with POS Support	0.98	0.98	0.98	0.98
POS-Morph Joint	0.96	0.96	0.96	0.96

(POS Results)				
POS-Morph Joint (Morph Results)	**0.98**	**0.98**	**0.98**	**0.98**

Table 5: Results for all proposed models

Feature	Accuracy	Precision	Recall	F1-Score
Gender	0.97	0.97	0.97	0.97
Number	0.97	0.97	0.97	0.97
Type	0.99	0.99	0.99	0.99
Person	0.99	0.99	0.99	0.99
Tense	0.99	0.99	0.99	0.99
Case	0.98	0.98	0.98	0.98
Aspect	0.98	0.98	0.98	0.98
Overall Morph	**0.98**	**0.98**	**0.98**	**0.98**

Table 6: Morphological Feature Wise Results for POS-Morph joint prediction

It is important to compare the results of the proposed model with the similar existing work. Table 7 shows the result comparison for the Gujarati POS tagging with baseline models. We observe that most of the previous work was carried out on very small dataset compared to the present work. To compare the results of the transformer model with similar neural architecture, we also test our POS tagging model and use it as one of the baseline systems. For the morphological analysers also, we use earlier rule based as well as Bi-LSTM based work as baseline model and compare the results in Table 8.

Reference Work	Approach	Dataset Size	Accuracy
[42] (Yagink et al.,2017)	Statistical	351 words	92.87
[25] (Patel et al.,2008)	CRF	30,000 sentences	89.90
[26] (Tailor et al., 2021)	Rule Based + LSTM	5600 sentences	91.10
[27] (Jobanputra et al., 2021)	LSTM	1000 sentences	95
Present Work	Bi-LSTM	30,000 sentences	90.50
Present Work	Transformer	30,000 sentences	96

Table 7: Comparison of POS tagging results with baseline models

Reference work	Approach	Dataset size	Accuracy
[28]	Paradigm based and statistical	500 words	92.34
[43]	Paradigm based	Not Known	82.16
[29]	Bi-LSTM based	16k words	91.67
Present work	**Transformer based**	**30000 sentences**	**98**

Table 8 : Comparison of Morph tagging results with baseline models

While it is evident from the result tables that the POS-Morph joint model gives the best results, we analyse the training data and observe that the loss convergence in the joint model is better than the standalone models. Figure 5 shows the comparison of loss between only morph model and the morph model with POS category as an input. We observe that the second model converges faster than the first model. This supports our argument that presence of POS category while predicting morph tag helps the model.

Sentence	Output of Only POS Model	Output of POS+MORPH model	Observation
સ્વામી રાજી થયા.	root=રાજી, POS=JJ, Type=LGSPEC01	root=રાજી, POS=JJ, Type=LGSPEC01	For the word રાજી, Both models generate correct output.
રાજી એ શાક ખાધું.	root=રાજી, POS=JJ, Type=LGSPEC01	root=રાજી, POS=NN, Gender=FEM, Case=NOM	For the word રાજી, POS+MORPH model generates correct output.
તું મોબાઈલ નીચે મૂક	root=મુક, POS=JJ, Type=LGSPEC01	root=મુકવું, POS=V_VM, Tense=PST, Aspect=LGSPEC01	For the word મૂક, POS+MORPH model generates correct output.
અકિરાને મૂક ભાષા આવડતી હતી.	root=મુક, POS=JJ, Type=LGSPEC01	root=મુક, POS=JJ, Type=LGSPEC01	For the word મૂક, both models generates correct output.

Table 9: Analysis of the predicted outputs

In Table 9, we highlight some sentences having word ambiguities along with their corresponding outputs. These examples help in understanding the scenarios in which the POS-Morph joint model generates better output than the standalone model. The example sentences are selected in such a way that the same word belongs to different POS category in two different sentences. For example, consider a word રાજી. This word is used as an adjective in the first sentence and as a proper noun in the second sentence. For the first sentence, both models generate correct POS and Morph tags but for the second sentence, only the joint POS-Morph model generates proper output. We make similar observation for the word મૂક which means `silent' when used as an adjective and `to put` when used as verb.For the better understanding of the outputs generated by all models, we show the output of a single sentence using all 4 models in Table 10.

Input Sentence	માછીમારે માછલીને મારી નાખી.
Transliteration	Māchīmārē māchalīnē mārī nākhī.
English Translation	The fisherman killed the fish.
Output of Model A (Only POS)	[('માછીમારે', {'pos': 'N_NN'}), ('માછલીને', {'pos': 'N_NN'}), ('મારી', {'pos': 'V_VM'}), ('નાખી', {'pos': 'V_VAUX'})]
Output of Model B (Only Morph)	[('માછીમારે', {}), ('માછલીને', {}), ('મારી', {'gender': 'FEM', 'type': 'LGSPEC02'}), ('નાખી', {'gender': 'FEM', 'number': 'PL', 'type': 'LGSPEC03', 'person': '3', 'tense': 'PST'})]
Output of Model C (Morph with POS support)	[('માછીમારે', {}), ('માછલીને', {}), ('મારી', {'gender': 'FEM', 'type': 'LGSPEC02'}), ('નાખી', {'gender': 'FEM', 'number': 'PL', 'type': 'LGSPEC03', 'person': '3', 'tense': 'PST'})]
Output of Model D (Joint POS-Morph)	[('માછીમારે', {'pos': 'N_NN'}), ('માછલીને', {'pos': 'N_NN'}), ('મારી', {'pos': 'V_VM', 'gender': 'FEM', 'type':

	'LGSPEC02'}), ('ગાખી', {'pos': 'V_VAUX', 'gender': 'FEM', 'number': 'PL', 'type': 'LGSPEC03', 'person': '3', 'tense': 'PST'})]

Table 10: Outputs of all models for a given sentence

6. Conclusion

In this work, we have developed efficient model for the prediction of POS category and morph features for Gujarati language. Our model of jointly predicting POS category and morphological features effectively captures the intricate relationship between POS-Morph features from the results, we conclude that the proposed model is best in terms of the dataset size and the results compared to all previous models. This work is a substantial contribution to the field of NLP for the Gujarati language as the proposed system can be used as an important component while building higher level NLP systems for the Gujarati language. Our experiments indicate that the joint model performs better than the standalone model in case of language ambiguities.

Author Contribution

First author is PhD student who has proposed the idea and carried out the survey and experiments. Second author is undergraduate student who has helped in programming and third author is research supervisor who has done overall supervision and helped in writing.

References

[1] G. Cardona and B. Suthar, "Gujarati," in *The Indo-Aryan languages*. Routledge, 2007, pp. 722–765.

[2] M. F. Porter, "An algorithm for suffix stripping," *Program*, 1980.

[3] K. Koskenniemi, "Two-level model for morphological analysis." in *IJCAI*, vol. 83, 1983, pp. 683–685.

[4] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Computational Linguistics 27(2)*, pp. 153– 198, 2005.

[5] M. Bapat, H. Gune, and P. Bhattacharyya, "A paradigm-based finite state morphological analyzer for marathi," in *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, 2010, pp. 26–34.

[6] A. Kumar, V. Dhanalakshmi, R. Rekha, K. Soman, S. Rajen- dran *et al.*, "Morphological analyzer for agglutinative languages using machine learning approaches," in *2009 International Conference on Advances in Recent Technologies in Commu- nication and Computing*. IEEE, 2009, pp. 433–435.

[7] D. K. Malladi and P. Mannem, "Context based statistical morphological analyzer and its effect on Hindi dependency parsing," *SPMRL 2013 - 4th Workshop on Statistical Parsing of Morphologically Rich Languages, Proceedings of the Workshop*, no. October, pp. 119–128, 2013.

[8] C. Malaviya, M. R. Gormley, and G. Neubig, "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)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2653–2663. [Online]. Available: https://www.aclweb.org/anthology/P18-1247

[9] G. Heigold, G. Neumann, and J. van Genabith, "An extensive empirical evaluation of character-based morphological tagging for 14 languages," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 505–513. [Online]. Available: https://www.aclweb.org/anthology/E17-1048

[10] D. Kondratyuk, "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*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 12–18. [Online]. Available: https://aclanthology.org/W19-4203

[11] P. Singh, G. Rutten, and E. Lefever, "A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek," in *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 128–137. [Online]

[12] E. C. Acikgoz, T. Chubakov, M. Kural, G. Şahin, and D. Yuret, "Transformers on multilingual clause-level morphology," in

Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 100–105. [Online].

Available: https://aclanthology.org/2022.mrl-1.10

[13] E. Brill, "A simple rule-based part of speech tagger," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[14] M. Divyapushpalakshmi and R. Ramalakshmi, "An efficient sentimental analysis using hybrid deep learning and optimiza- tion technique for twitter using parts of speech (pos) tagging," *International Journal of Speech Technology*, vol. 24, pp. 329– 339, 2021.

[15] B. Pham, "Parts of speech tagging: Rule-based," 2020.

[16] M. Constant and A. Sigogne, "Mwu-aware part-of-speech tag- ging with a crf model and lexical resources," in *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world*, 2011, pp. 49–56.

[17] T. D. Singh, A. Ekbal, and S. Bandyopadhyay, "Manipuri pos tagging using crf and svm: A language independent approach," in *proceeding of 6th International conference on Natural Language Processing (ICON-2008)*, 2008, pp. 240–245.

[18] T. Dalai, T. K. Mishra, and P. K. Sa, "Part-of-speech tagging of odia language using statistical and deep learning based approaches," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 6, pp. 1–24, 2023.

[19] R. D. Deshmukh and A. Kiwelekar, "Deep learning techniques for part of speech tagging by natural language processing," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2020, pp. 76–81.

[20] A. Singh, C. Verma, S. Seal, and V. Singh, "Development of part of speech tagger using deep learning," *Int J Eng Adv Technol*, vol. 9, no. 1, pp. 3384–91, 2019.

[21] P. Srivastava, K. Chauhan, D. Aggarwal, A. Shukla, J. Dhar, and V. P. Jain, "Deep learning based unsupervised pos tagging for sanskrit," in *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 2018, pp. 1–6.

[22] A. A. Maksutov, V. I. Zamyatovskiy, V. O. Morozov, and S. O. Dmitriev, "The transformer neural network architecture for part-of-speech tagging," in *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*. IEEE, 2021, pp. 536–540.

[23] H. Li, H. Mao, and J. Wang, "Part-of-speech tagging with rule- based data preprocessing and transformer," *Electronics*, vol. 11, no. 1, p. 56, 2021.

[24] M. V. Gamit, R. Joshi, and E. Patel, "A review on part-of- speech tagging on gujarati language," *International Research Journal of Engineering and Technology (IRJET)*, 2019.

[25] C. Patel and K. Gali, "Part-of-speech tagging for gujarati using conditional random fields," in *Proceedings of the IJCNLP-08 workshop on NLP for less privileged languages*, 2008.

[26] C. Tailor and B. Patel, "Hybrid pos tagger for gujarati text," in *Soft Computing and its Engineering Applications: Second International Conference, icSoftComp 2020, Changa, Anand, India, December 11–12, 2020, Proceedings 2*. Springer, 2021, pp. 134–144.

[27] C. Jobanputra, N. Parikh, V. Vora, and S. K. Bharti, "Parts- of-speech tagger for gujarati language using long-short-term- memory," in *2021 International Conference on Artificial Intel- ligence and Machine Vision (AIMV)*. IEEE, 2021, pp. 1–5.

[28] J. Baxi, P. Patel, and B. Bhatt, "Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods," *Proceedings of the 12th International Conference on Natural Language Processing*, no. December, pp. 178–182, 2015. [Online]. Available: https://www.aclweb.org/anthology/W15-5927

[29] J. Baxi and B. Bhatt, "Morpheme boundary detection & gram- matical feature prediction for gujarati : Dataset & model," in *Proceedings of the 18th International Conference on Natural Language Processing*, NIT, Silchar, Dec. 2021.

[30] ——, "A bidirectional-lstm based morphological analyzer for gujarati," In Press, (In Press).

[31] J. Baxi and b. bhatt, "Gujmorph - a dataset for creating gujarati morphological analyzer," in *Proceedings of the Language*

Resources and Evaluation Conference. Marseille, France: European Language Resources Asso- ciation, June 2022, pp. 7088–7095. [Online]. Available: https://aclanthology.org/2022.lrec-1.767

[32] M. Straka, J. Hajič, and J. Straková, "UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4290–4297. [Online]. Available: https://aclanthology.org/L16-1680

[33] M. Straka, J. Straková, and J. Hajic, "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*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 95–103. [Online]. Available: https://aclanthology.org/W19- 4212

[34] *K. Batsuren, O. Goldman, S. Khalifa, N. Habash, W. Kieraś, G. Bella, B. Leonard, G. Nicolai, K. Gorman, Y. G. Ate, M. Ryskina, S. Mielke, E. Budianskaya, C. El- Khaissi, T. Pimentel, M. Gasser, W. A. Lane, M. Raj, M. Coler, J. R. M. Samame, D. S. Camaiteri, E. Z. Rojas, D. López Francis, A. Oncevay, J. López Bautista, G. C. S. Villegas, L. T. Hennigen, A. Ek, D. Guriel, P. Dirix, J.-P. Bernardy, A. Scherbakov, A. Bayyr-ool, A. Anastasopoulos, R. Zariquiey, K. Sheifer, S. Ganieva, H. Cruz, R. Karahóǧa, S. Markantonatou, G. Pavlidis, M. Plugaryov, E. Klyachko, A. Salehi, C. Angulo, J. Baxi, A. Krizhanovsky, N. Krizhanovskaya, E. Salesky, C. Vania, S. Ivanova, J. White, R. H. Maudslay, J. Valvoda, R. Zmigrod, P. Czarnowska, I. Nikkarinen, A. Salchak, B. Bhatt, C. Straughn, Z. Liu, J. N. Washington, Y. Pinter, D. Ataman, M. Wolinski, T. Suhardijanto, A. Yablonskaya, N. Stoehr, H. Dolatian, Z. Nuriah, S. Ratan, F. M. Tyers, E. M. Ponti, G. Aiton, A. Arora, R. J. Hatcher, R. Kumar, J. Young, D. Rodionova, A. Yemelina, T. Andrushko, I. Marchenko, P. Mashkovtseva, A. Serova, E. Prud'hommeaux, M. Nepomniashchaya, F. Giunchiglia, E. Chodroff, M. Hulden, M. Silfverberg, A. D. McCarthy, D. Yarowsky, R. Cotterell, R. Tsarfaty, and E. Vylomova, "UniMorph 4.0: Universal Morphology," in Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, Jun. 2022, pp. 840– 855. [Online]. Available: https://aclanthology.org/2022.lrec- 1.89*

[35] *J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal dependencies v1: A multilingual treebank collection," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 1659–1666. [Online]. Available: https://www.aclweb.org/anthology/L16-1262*

[36] O. Goldman and R. Tsarfaty, "Morphology without borders: Clause-level morphology," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1455–1472, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.83

[37] M. Hämäläinen, N. Partanen, J. Rueter, and K. Alnajjar, "Neural morphology dataset and models for multiple languages, from the large to the endangered," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, May 31–2 Jun. 2021, pp. 166–177. [Online]. Available: https://aclanthology.org/2021.nodalida-main.17

[38] K. Batsuren, G. Bella, and F. Giunchiglia, "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*, G. Nicolai, K. Gorman, and R. Cotterell, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 39–48. [Online]. Available: https://aclanthology.org/2021.sigmorphon-1.5

[39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[40] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhat- tacharyya, M. M. Khapra, and

P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Find- ings of EMNLP*, 2020.

[41] R. Joshi, "L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages," *arXiv preprint arXiv:2211.11418*, 2022.

[42] A. Yajnik and M. Prajapati, "Part of speech tagging using statistical approach for gujarati text," *Int J Appl Res Sci Eng*, 2017.

[43] D. N. Shah and H. Bhadka, "Paradigm-based morphological an- alyzer for the gujarati language," in *Intelligent Communication, Control and Devices: Proceedings of ICICCD 2018*. Springer, 2020, pp. 469–481.