International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

A survey on Morphological Feature Extraction for Gujarati

Jeenal Patel[1], Dr. Pooja Bhatt2
[1]Research Scholar, Parul University, 391760, India
[2]Associate Professor, Parul University, 391760, India
Email ID:*2423004700013@paruluniversity.ac.in[1], pooja.bhatt28403@paruluniversity.ac.in[2]*

Abstract
The field of linguistics known as morphology focuses on the smallest meaningful units, known as morphemes, and examines the internal structure and formation of words. It studies how prefixes, suffixes, and roots work together to form new words, shift grammatical categories (such as noun to adjective), or modify number and tense. The following datasets are used in this paper for morphological analysis in the Gujarati language: Gujmorph, TDIL-ILCI-II corpus, Rudhiprayog ane kahevatsangrah , Gujarati Lexicon, and EMILLE corpus. In this review on a bidirectional LSTM-based morphological analyzer for Gujarati, the authors show that across important POS categories, the Bi-LSTM with Individual Label Representation method outperforms the Bi-LSTM monolithic and Bi-LSTM individual feature representation approaches in terms of accuracy. The accuracy increased from 68.27% (unsupervised) and 70.64% (individual feature representation) to 99.95% for nouns, from 12.95% and 16.18% to 78.76% for verbs, and from 25.72% and 85.85% to 99.84% for adjectives. Dataset Expansion: Future research should focus on making the current training datasets larger and include other POS categories outside of nouns, verbs, and adjectives.India.

Keywords: *Computational Morphology, Gujarati Language, GujMORPH, Grammatical Feature Prediction, Bi-Directional LSTM, Gujarati-BERT, Hybrid Morphological Analyzer.*

1. Introduction
In Natural Language Processing (NLP), speech (POS) tagging and morphological analysis are basic word-level operations that are very important for understanding the grammatical structure and semantic meaning of a language[1]. A very useful linguistic resource is a morphological analyzer, which decomposes an inflected word into its constituent morphemes (root and suffixes) and identifies the grammatical categories associated with them[2]. The effectiveness of various NLP applications, such as text summarization, machine translation, sentiment analysis, and question answering, relies on these resources. There are specific challenges in developing these resources for the Gujarati language. Gujarati is the 26th most widely spoken Indo-Aryan language in the world, with 55 to 65 million speakers[3]. It is classified as a low-resource language despite having a large number of speakers because of there are no computational resources available, such as dictionaries and corpora[4]. Gujarati has a complex morphology and a large number of inflections. Unlike other Indo-Aryan languages such as Hindi, which have only two genders and Gujarati has three genders (masculine, feminine, and neuter) and a Subject-Object-Verb (SOV) word order[5]. The conjugation of the Gujarati verb is quite complex, as it depends on the tense, aspect, mood, and person, while the inflection of the noun depends on the gender, number, and case. In addition, the agglutinative nature of the language sometimes causes the merger of several morphemes into a single form of a word, which may cause ambiguity when a word can symbolize different morphological properties or POS tags depending on the context[6]. In the past, machine learning and deep learning models have been used in place of rule-based models for morphological analyzer construction. Rule-based models, such as paradigm-based models or finite-state transducers, require manual construction of suffix tables and word formation rules[7]. These models are difficult to maintain manually and often face ambiguities in to the natural language. Although they involved heavy feature engineering, traditional machine learning models

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

such as Support Vector Machines (SVM) and Conditional Random Fields (CRF) were also considered[8]. Bi-directional Long Short-Term Memory (Bi-LSTM) networks are the models enabled by current advancements in the deep learning. These models efficiently utilize past and future information for morpheme boundary recognition and grammatical feature tagging, eliminating the necessity for rule-based systems[9]. The application of pre-trained models using the Transformer architecture, such as BERT and its variants (mBERT, IndicBERT), has also been widely accepted in recent years. These models make it possible to predict POS tags and morphological features simultaneously to handle ambiguities in context[10]. The application of the high-quality data is essential to these developments. The GujMORPH dataset, which is represented in the standard Unimorph format, has been developed, which is a major achievement in Gujarati NLP[11]. The GujMORPH dataset, providing morpheme segmentation and feature annotation, enables the integration of Gujarati into global linguistic models. The collection has more than 16,000 distinct inflected forms. This overview illustrates the evolution of these computer applications and linguistic intricacy of Gujarati, which are two primary catalysts for these advancements.

2. Literature review

The research offers various methods for dealing with Gujarati morphology, script recognition, and idioms, both rule-based and machine learning. A research work proposes a novel morphological analysis-driven system for dynamically identifying inflected Gujarati idioms. The authors created a database of 3,410 distinct idioms, which grew to 6,047 valid inflected forms based on suffixes. With 43 linguistically defined rules, the system searches for possible inflections of base idioms and matches them with the input text. In the evaluation, the system correctly identified all 7,400 idiomatic forms in the test set. While traditional loss values were not shown, an important drawback was mentioned: idioms not in the database cannot be identified, resulting in lower coverage[6]. Another work is on a rule-based morphological stemmer for Gujarati, aiming to cope with the language's highly inflected morphology. The system uses 52 frequent inflections and follows a strategic rule set after text processing and character replacement. The non-iterative approach ensures low computational complexity. Tested on the EMILLE corpus with 20 news articles, the stemmer showed an accuracy of 92.41%. The errors were mostly due to over-stemming and under-stemming, which affected 61 out of 777 inflected words, with further inconsistency shown in processing loanwords[14].

Another research work is on the semi-automatic rule generation process for suffix replacement rules in Gujarati and Hindi morphological analysis. By comparing the root form dictionary with the inflected word corpora, the system filters and generates rules based on their impact. The Gujarati experiment on the EMILLE corpus resulted in a precision of 0.83 and an F-measure of 0.76. Performance degradation was observed when there were incorrect or absent root form entries in the dictionary[7]. Deep learning techniques are also investigated in a Bi-LSTM-based morphological analyzer for Gujarati. This analyzer not only detects morpheme boundaries but also assigns grammatical features. The most important aspect of this work is the prediction of grammatical features such as gender, number, and case as separate labels, as opposed to predicting them as classes, thus reducing the complexity of the output.The model was trained on a premier UniMorph dataset and comprising over 16,000 inflected words and attaining near-perfect accuracy in nouns and adjectives or alongside a significant enhancement in verb accuracy. Binary cross-entropy loss was employed, and minimal validation loss values were maintained to ensure effective generalization and reduced overfitting[2]. In another area, the authors worked on the offline Gujarati handwritten OCR problem, focusing on characters that do not have the Devanagri shirorekha. As there were no available benchmarks, a new benchmark was established with the assistance of more than 100 writers. The approach utilized Freeman chain code for feature extraction and Hidden Markov Models for classification. While neural networks were more accurate and the chain code approach offered lossless compression and simplicity of implementation[5].

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

The development of GujMORPH, a large-scale annotated Gujarati morphological corpus, is also discussed. Based on the ILCI-II corpus and UniMorph framework, the corpus was developed for segmentation and grammatical tagging. The baseline Bi-LSTM model achieved 89.05% accuracy on the morpheme boundary detection, although verb feature prediction was difficult because of intricate inflection patterns[3]. In addition, a hybrid morphological analyzer that integrates rule-based models, dictionary search, and statistical disambiguation was also discussed. On the gold standard set of 500 words, the hybrid approach reached over 92% accuracy when applying linguistic knowledge, but had difficulty with the lack of dictionary information and derivational morphology[15]. Finally, the transformer-based models, such as Gujarati BERT, were fine-tuned for the joint task of POS tagging and morphological feature prediction. With a combined ILCI-II and UniMorph dataset of 30,000 sentences, the joint model resulted in F1-scores of 0.98 for morphology and 0.96 for POS tagging. Analysis of training indicated stable convergence of loss values, with joint learning helping to improve the understanding of context and alleviate ambiguity[5]. There is considerable work on morphological analyzers and stemmers for other languages too. Conventional approaches, such as Porter's stemmer, are primarily based on manually designed suffix replacement rules. In work on derivational stemming involving prefixes, Hull (1996) and Krovetz (1993) contributed to this area, but Hull cautioned against the dangers of over-stemming. The first lightweight stemmer for Hindi in Indian languages was proposed by Ramanathan and Rao (2003). Subsequently, Shrivastava et al. (2005) designed a high-precision stemmer using 86 manual criteria. More recently, tools such as YASS and research by Pandey and Siddiqui (2008) have concentrated on unsupervised, corpus-based learning to automatically generate rules[3]. Gujarati is a low-resource, morphologically rich language that is very challenging for NLP. From the literature available, there has been a transition from statistical approaches such as HMM and CRF, and rule-based stemmers to deep learning approaches such as Bi-LSTMs. Although these approaches were more efficient, they sometimes lacked context or had to resort to manual feature extraction. It is important to note that morphological analysis and POS tagging are normally treated as two distinct tasks in previous research. In the case of Gujarati, transformer- based approaches such as BERT have not been adequately investigated, although they are very popular in other languages. This paper will fill this gap by proposing a collaborative prediction approach using Gujarati-BERT based on the linguistic dependency between features[15]. Unlike other forms of the print media, this survey highlights the lack of work on the challenges posed by offline handwritten OCR for the Gujarati script. To overcome the lack of benchmark data, the authors propose a new dataset that has been gathered from more than 100 writers. Different methods of the feature extraction, such as Moments, Zoning, and Freeman chain code, which allows for lossless storage are analyzed. The accuracy compromises of models such as SVM and neural networks are also analyzed for classification. The paper concludes with the recommendation of Hidden Markov Models as a good, faster probabilistic method for character recognition[4].

3. Comparative Analysis

TABLE 1

Sr.NO	Title	Author Name	Method	Accuracy	Loss	Future Scope
1	GujMORPH - A Dataset for Creating Gujarati Morphological Analyzer[6]	Jatayu Baxi, Brijesh Bhatt	Bi-LSTM baseline system.	89.05% (boundary detection); F1: 0.68 (Noun), 0.12 (Verb), 0.68 (Adj).	Not explicitly specified in this excerpt.	Expand the dataset with more examples and include remaining POS categories.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

2	Morpheme Boundary Detection & Grammatical Feature Prediction for Gujarati : Dataset & Model[14]	Jatayu Baxi, Brijesh Bhatt	Bi-Directional LSTM using binary encoding for segmentation.	89.05% (segmentation) ; Tagging: 70.64% (Noun), 16.18% (Verb), 85.85% (Adj).	Binary cross entropy (segmentation) and categorical cross entropy (tagging).	Implement seq2seq models and study sentence-level dependency for analysis.
3	Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods[7]	Jatayu Baxi, Pooja Patel, Brijesh Bhatt	Hybrid approach: Paradigm-based + Wordnet + Statistical probability.	92.34% (Knowledge-based hybrid) and 82.84% (Statistical hybrid).	7.66%	Address derivational morphology, which the current system cannot handle.
4	A Stemmer for morphological level analysis of Gujarati language[2]	Jikitsha Sheth, Bankim Patel	Rule-based stemming using 52 common inflections and substitution rules.	92.41%.	Not explicitly specified.	Enrich the rule set to minimize overstemming/understemming and apply a hybrid statistical approach
5	A bidirectional LSTM-based morphological analyzer for Gujarati (2024 Edition)[5]	Jatayu Baxi, Brijesh Bhatt	Bi-LSTM with Individual Label Representation .	Noun: 99.95%, Verb: 78.76%, Adjective: 99.84%.	Training Loss: 0.096, Validation Loss: 0.079 (Binary Cross Entropy).	Predict POS and morphological features side-by-side to study their mutual effect
6	Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages[3]	Niraj Aswani, Robert Gaizauskas	Rule-based system with semi-automatically acquired suffix rules.	(Gujarati) Precision: 0.83, Recall: 0.70, F-measure: 0.76.	Precision:0.17	Improve the Gujarati base-form list (GRFList) to enhance system accuracy

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

7	Part of Speech and Morph Category Prediction for Gujarati[15]	Jatayu Baxi, Om Soni, Brijesh Bhatt	Transformer-based pre-trained model (GujaratiBERT) with joint prediction.	Joint Morph F1: 0.98, Joint POS F1: 0.96.	Joint Training Loss: 86.15; Standalone POS Training Loss: 22.13.	Use as a core component for higher-level NLP like machine translation for Gujarati
8	Developing a Hybrid Morphological Analyzer for Low-Resource Languages[13]	Musica Supriya, D. Acharya Udupi, et al.	Hybrid: Rule-based (Apertium Lttoolbox) + Transformer.	Precision: 0.924, Recall: 0.925, F1 score: 0.925.	CrossEntropyLoss (used during 10,000 training steps).	Expand onto verbal inflections and handle compound/derivational words
9	A Novel Morphological Analysis based Approach for Dynamic Detection of Inflected Gujarati Idioms[1]	Jatin C. Modh, Jatinderkumar R. Saini, and Ketan Kotecha	A rule-based morphological analysis approach. The method involves identifying base/stem forms of idioms and applying a reverse rules generation process to create 43 specific rules for generating inflected forms. The algorithm follows a process of UTF-8 text input, preprocessing (whitespace and special character removal), tokenization, and dynamic idiom	The proposed model successfully detected all static and inflected Gujarati idiom forms present in the tested input text. The testing involved 7400 different idiom forms, and the correctness was verified by two linguists with doctorate degrees in the Gujarati language.	The sources do not provide a numerical "loss" metric. This metric is common in machine learning models but is not mentioned for this specific rule-based morphological system.	Future work includes collecting all Gujarati idioms from all possible sources to expand the current database and overcome the limitation of the system only recognizing idioms already stored. Additionally, researchers aim to implement the model in real-world machine translation to improve the translation of Gujarati idiomatic text into other languages.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

			detection by comparing input text against the rule-generated forms and an idiom database.			
10	A Survey on Gujarati Handwritten OCR using Morphological Analysis[4]	Vaidehi Patel and Prof. Abhinay Pandya	The paper focuses on **offline handwriting recognition** using **Freeman chain code** for feature extraction and the **Hidden Markov Model (HMM)** for classification. It also surveys other methods such as Zoning, Moments, Neural Networks, and SVM	Although the sources don't give a precise numerical accuracy for the authors' implementation, they do mention that SVMs offer "high accuracy" and neural networks have a "higher recognizing ratio"	**SVM:** Known for having **"poor recall"** despite high accuracy	Future objectives include the advancement of more efficient OCR systems utilizing Hidden Markov Models, the creation of benchmark datasets, and the identification of suitable segmentation algorithms and diacritic management for the Gujarati language.

4. Result

This evaluation of the extraction of Gujarati morphological properties indicates a marked as transition from the rule-based systems to more complex hybrid approaches and deep learning that resulting in significant improvements in the strength and accuracy of the models. The first rule-based stemmers and paradigm-based analyzers offered a strong language platform that was accurate to about 92%. However, they had limitations in terms of scope, the possibility of a missing dictionary definition, and the complexity of dealing with ambiguity. The employment of structured data sources, such as GujMORPH, facilitated a much faster progress since it was easier to carry out a comprehensive evaluation. The Bi-LSTM models indicated significant improvements in morpheme boundary detection and grammatical property prediction.The Bi-LSTM model with separate label representation showed near-perfect results for some POS tags, as described in A bidirectional LSTM-based morphological analyzer for Gujarati, with noun and adjective accuracy rates above 99% and significant improvements in verb prediction. More recent transformer-based models like Gujarati-BERT pushed the boundaries further by allowing simultaneous POS and morphological prediction, with F1-scores of 0.98 for morphology and 0.96 for POS tagging. These results clearly establish the efficacy of contextual modeling and joint learning in alleviating the ambiguity of morphologically rich Gujarati text. Hybrid morphological analyzers that integrate rule-based approaches with neural or transformer modules showed balanced performance,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

especially in dealing with out-of-vocabulary and low-resource scenarios. A comparison of the results suggests that hybrid morphological analyzers always score F1-scores of 0.92-0.93, providing a good compromise between linguistic interpretability and statistical generalizability.

Conclusion

For improving NLP in low-resource languages like Gujarati and there is a requirement for high-quality computational morphology resources. According to a study, the creation of the supervised models needs high-quality datasets like GujMORPH, which are UniMorph annotated. Regarding the creation of models, although rule-based models face issues while dealing with ambiguity, Bi-LSTM and the Transformer model, a deep learning model, can easily cope with complex patterns without human assistance. The application of a optimization method with varying representations of labels is a major advancement in terms of accuracy because it decreases the number of classes in the output. Moreover, the application of POS-Morph models that include Gujarati-BERT models is more efficient in handling linguistic ambiguity rather than applying models separately because of the correlations among the features. Finally, for the Kannada language, hybrid models that include rule- based models and Transformers are extremely efficient in handling out-of-vocabulary words.

Future Scope

Although considerable progress has been achieved, there are still some areas that could be explored in future research to further improve the performance and usefulness of morphological analyzers. Dataset Expansion: Future research should target the expansion of the current training datasets and the incorporation of the remaining categories of POS tags, apart from nouns, verbs, and adjectives. Complex Linguistic Structures: The majority of the current studies are based on inflectional morphology. Future studies should target derivational morphology, compound words, and the handling of ancient language variants or foreign loan words.Downstream Integration: The next important step would be the integration of morphological analyzers into higher-level NLP tasks, such as machine translation , question answering , sentiment analysis , and text summarization , to assess their effect on the performance of the overall system.

References

[1]. Modh, Jatin C., Jatinderkumar R. Saini, and Ketan Kotecha. "A Novel Morphological Analysis based Approach for Dynamic Detection of Inflected Gujarati Idioms."International Journal of Advanced Computer Science and Applications 13.4 (2022).

[2]. Sheth, Jikitsha, and Bankim Patel. "Dhiya: A stemmer for morphological level analysis of Gujarati language." 2014 international conference on issues and challenges in intelligent computing techniques (ICICT). IEEE, 2014.

[3]. Aswani, Niraj, and Robert J. Gaizauskas. "Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages." LREC.2010.

[4]. Patel, Vaidehi, and Abhinay Pandya. "A Survey on Gujarati Handwritten OCR using Morphological Analysis." (2016): 2395-1990.

[5]. Baxi, Jatayu, and Brijesh Bhatt. "A bidirectional LSTM-based morphological analyzer for Gujarati." Natural Language Processing 31.2 (2025): 198-214.

[6]. Baxi, J., and B. Bhatt. "GujMORPHADatasetforCreatingGujaratiMorphological Analyzer." ProceedingsoftheThirteenthLanguageResourcesandE valuationConference (2022).

[7]. . Baxi, Jatayu, Pooja Patel, and Brijesh 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. 2015.

[8]. Patel, Chirag, and Karthik Gali. "Part-of-speech tagging for Gujarati using conditional random fields."

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0115
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 749 - 756

Proceedings of the IJCNLP-08 workshop on NLP for less privileged languages. 2008.
[9]. Dua, Mohit, et al. "A review on Gujarati language based automatic speech recognition (ASR) systems." International Journal of Speech Technology 27.1 (2024): 133-156.
[10]. Baxi, Jatayu, and Brijesh Bhatt. "Recent advancements in computational morphology: A comprehensive survey." arXiv preprint arXiv:2406.05424 (2024).
[11]. Panchal, Brijeshkumar Y., and Apurva Shah. "A Survey on Gujarati NLP Research Work." (2025).
[12]. Baxi, Jatayu, and Brijesh Bhatt. "GujMORPH-A dataset for creating Gujarati morphological analyzer." Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.
[13]. Supriya, Musica, et al. "Developing a Hybrid Morphological Analyzer for Low-Resource Languages." Applied Sciences 15.10 (2025): 5682.
[14]. Baxi, Jatayu, and Brijesh Bhatt. "Morpheme boundary detection & grammatical feature prediction for Gujarati: Dataset & model." arXiv preprint arXiv:2112.09860 (2021).
[15]. Baxi, Jatayu, Om Soni, and Brijesh Bhatt. "INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING."