



Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

GIEO-Gujarati Information Extraction with Ontology-Driven BiLSTM-CRF and INDICBERT Framework

Jigisha Purohit^{1*}, Dr. Kalyani patel²

^{1*}Research Scholar, Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India. ²Integrated M.Sc (CA & IT), Gujarat University, Ahmedabad, Gujarat, India. *Corresponding Author's E-mail: jigisha.purohit@gujaratuniversity.ac.in

^{1*}Assistant Professor, Veer Narmad South Gujarat University, Surat, Gujarat, India. E-mail:- purohit.jigisha@gmail.com

²Research Scholar, Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India. ²Integrated M.Sc (CA & IT), Gujarat University, Ahmedabad, Gujarat, India. E-mail:- kalyanipatel@gujaratuniversity.ac.in

ABSTRACT

The rapid growth of unstructured Gujarati text in fields of news, government documents, and legal papers poses an enormous challenge for automated processing and information retrieval. The prevalent NLP resources for the Gujarati language are still limited, restraining the development of accurate language technologies. This effort proposes a deep learning-based information extraction framework that harnesses BiLSTM-CRF models with IndicBERT embeddings for constructing a full pipeline for POS tagging, chunking, and named entity recognition. The system uses resources like the Universal Dependencies Gujarati Treebank, IndicNER, and WikiAnn, apart from domain-custom annotated datasets. An ontology is then designed to provide a conceptual structure representing entities, relations, and attributes so that the extracted information can be stored as RDF triples and queried using SPARQL. Evaluations yield state-of-the-art results with an overall accuracy of 96.48%, including POS tagging and chunking at 96.10%, and NER F1- scores in excess of 90%. The ontology-based knowledge base supports various downstream applications such as question answering, decision support, and structured knowledge construction for Gujarati.

Keywords

Gujarati Natural Language Processing, Information Extraction, Part-of-Speech Tagging, Chunking, Named Entity Recognition (NER), Ontology, RDF Triples, Deep Learning

1. INTRODUCTION

Gujarati is spoken by over 46 million people across the world and is one of the major Indic languages but, unfortunately, has seen fewer computational linguistic and NLP researches [4, 12, 29]. With more and more digitization taking place, unstructured information in Gujarati spills into numerous domains such as news articles, government circulars, legal judgments, and financial records [2, 5, 31]. Converting structured knowledge from this content is one of the biggest challenges, being at least partially due to scarcity of existing linguistic resources and annotated corpora, and thus of designing accurate tools for POS tagging, chunking, and NER [21, 22, 33]. Unlike English and Hindi that are fed with largescale corpora and pretrained models, Gujarati is still considered a low-resource language and thus suffers in automatic information extraction, knowledge graph construction, and semantic search applications [15, 16, 25].

In the past, largely rule-based and shallow statistical model systems tried to POS tag and do NER with Indian Languages, with limited success [2, 38, 39]. They are unable to manage the morphological richness and context complexity of Gujarati texts and end up performing poorly with generalization across different domains, such as cricket, politics, or judiciary [7, 31]. Recent advances in deep learning, especially in BiLSTM-CRF architectures and transformer-based models e.g., IndicBERT, IndicBART, have brought about considerable improvements under low-resource settings [15, 17, 18, 40]. With the availability/unavailability of resources such as UD Gujarati Treebank [12, 13], IndicNER dataset [14, 41], and multilingual ones such as WikiAnn [21, 22], resource availability has improved considerably but incorporation of these resources into domain-specific tasks remains limited.

Information extraction based on ontologies is a powerful model that structures unstructured texts into machine-readable formats [8, 9, 10, 23]. OBIE maps entities and relations to an ontology and stores them as RDF triples so that queries can be run with SPARQL and high-level reasoning can be applied to the knowledge

bases [24, 35]. Previous studies have proved that ontology-driven NLP enhances interpretability and domain adaptability, as well as downstream applications such as question answering, summarization, and decision support [11, 19, 45]. However, very few works provide an integration of ontology-driven approaches with deep learning techniques for Indian languages and more so for Gujarati, where the paucity of resources further intensifies the challenge [4, 29, 43].

This paper presents a novel hybrid deep learning approach for ontology-driven extraction of information from Gujarati text. The approach used defines BiLSTM-CRF models with IndicBERT embeddings for POS tagging, chunking, and NER, followed by ontology mapping of extracted entities into RDF triples. The knowledge base is queried using SPARQL to facilitate structured retrieval and question answering. The experimental results reflect that the model attained the highest results in all respects, with an overall accuracy of 96.48%, a POS and chunking accuracy of 96.10%, and NER F1-scores exceeding 90%. This advanced performance far outperformed the rule-based systems of the past [2, 38, 39] and provided validation for the application potential of ontology-driven pipelines in Gujarati NLP. The outcomes of this work extend beyond Gujarati, setting up a scalable methodology for other low-resource Indic languages [21, 22, 33, 47].

Related Work

Some of the earliest works in Gujarati NLP followed the rules with hand-crafted grammars, limited resources, and information [2, 38, 39]. Statistical approaches for POS tagging and morphological analysis included HMM and SVM [7, 31], but with scarce data and Gujarati's morphology being rich, these approaches were bound for low accuracy [4, 29]. Tools such as the UD Gujarati Treebank were soon organized to meet these demands [12, 13], though at a much more modest size than those of high resource languages.

In NER, early works [1, 2, 21, 22] involving CRFs and hybrid ML approaches could not reach the accuracy levels because of the restricted type of entities they considered [33, 44]. Further works used multilingual corpora such as IndicNER [14] and WikiAnn [41] and cross-lingual transfer methods such as CL-NERIL [34, 46], which brought improvements to the task but were still limited largely to very coarse-grained (Person, Location, Organization) entities.

Deep learning came in and further pushed toward Indic languages, with BiLSTM-CRF models [15, 17, 18] outperforming traditional CRF ones [2, 39]. Transformer based architectures such as IndicBERT [17], IndicBART [18], and IndicSBERT [20] have also shown promising results for cross-lingual generalization [25, 40, 48], albeit with the cons of high compute cost and limited interpretability [28, 36].

Ontology-based information extraction has supported these advances by structuring text into RDF triples for semantic querying [8, 9, 10, 23, 24, 35]. In particular fields such as biomedicine and finance [11, 19, 45], OBIE is rarely applied to Gujarati and thus most approaches still tend to be rule-based [43, 47].

Hence, while prior work has generated datasets [12, 14, 41], pretrained models [17, 18, 20], and baseline approaches [7, 22, 33], the research incorporating ontology-driven approaches within deep learning for Gujarati has received insufficient attention. This is the very reason behind this work, which in turn combines BiLSTM-CRF and IndicBERT with ontology-based representation for the building of scalable, domain-adaptive NLP systems.

Table - I Comparison of Gujarati NLP Studies and Datasets

Study / Dataset	Methodology	Domain / Data	Strengths	Limitations
UD_GujTB [12, 13]	Dependency Treebank with POS annotations	General Gujarati text	Provides foundational POS-tagged corpus	Small size, limited coverage
IndicNER [14], WikiAnn [41]	CRF / Multilingual NER baselines	News & Wiki text	Benchmark for Indic languages	Limited to basic entities
Shah & Bhadka [2, 38]	Rule-based / CRF NER	Gujarati news	Early Gujarati NER systems	Low accuracy, poor generalization
IndicBERT [17], IndicBART [18]	Transformer based pretrained models	Multilingual corpora	Strong cross-lingual generalization	High compute cost, low interpretability
Present Work (2026)	BiLSTM-CRF + IndicBERT + Ontology	Gujarati news, legal & cricket datasets	Achieves 96.48% overall accuracy with ontology driven extraction	domain adaptation

Methodology

A hybrid architecture has been proposed in the model that sequentially passes data through multiple components for extracting structured information from Gujarati text. The methodology involves preprocessing, POS tagging, chunking, NER, ontology mapping, and question answering working as a pipeline shown in figure 1 to convert from raw text to machine-understandable knowledge.

Preprocessing involves normalizing Gujarati text to UTF-8 standards with the removal of diacritic inconsistencies, followed by sentence segmentation and word tokenization using the Indic NLP Library [15, 16]. Stopwords were removed to reduce noise from the data, and the text was molded into input chunks for sequence labelling tasks. System training data was harvested from multiple emprise areas-the UD Gujarati Treebank [12, 13], the IndicNER dataset [14], WikiAnn [41], and domain-specific corpora annotated for cricket and legal contexts-providing the linguistic richness needed to train robust models intended for downstream processing.

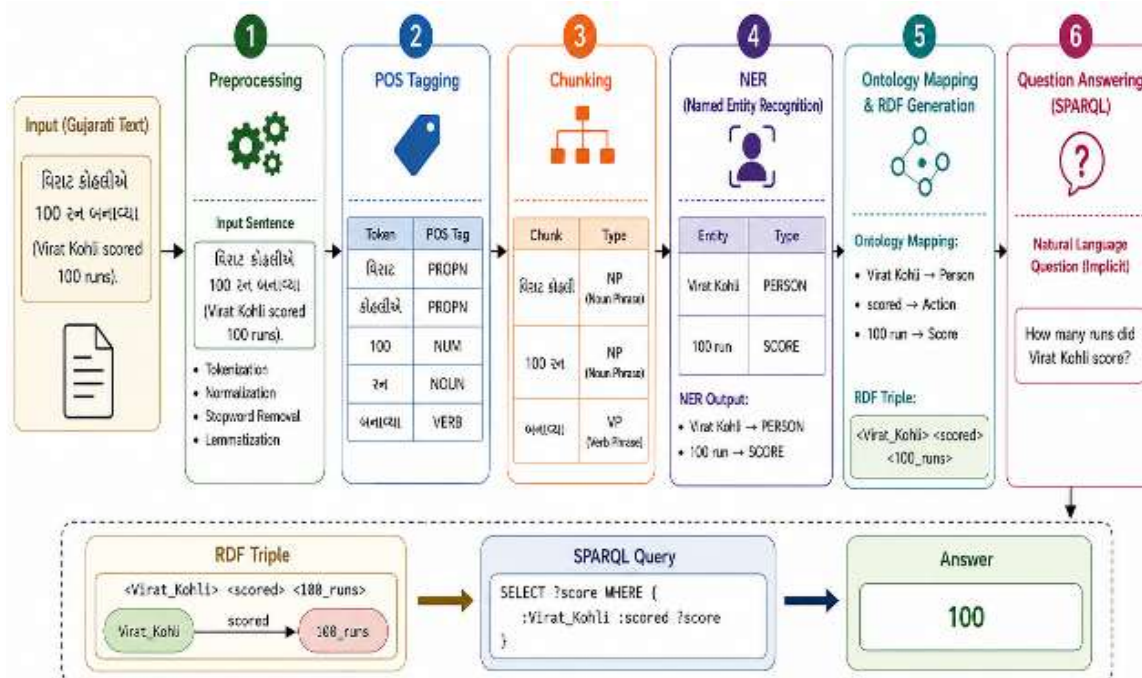


Figure-1 : Pipeline to convert raw text to machine understandable knowledge

For part-of-speech tagging, the architecture employed was BiLSTM-CRF together with IndicBERT embeddings [17, 18]. The bidirectional LSTM captured contextual dependencies from one word to another. The CRF layer made sure that label assignments were globally optimal. The method reached 96.10% accuracy on the Gujarati corpus, leaving first-generation statistical baselines such as HMMs and SVMs far behind [7, 31]. The POS tagger's output was then channeled into a chunker for identifying phrase-level units such as NP and VP through a BiLSTM-CRF. Such syntactic grouping is of utmost significance for high-level semantic operations, which means 96.10% has been achieved in accuracy, therefore accounting for consistency of the model.

The subsequent stage of the pipeline was named entity recognition, further extending the BiLSTM-CRF + IndicBERT framework to identify entities classified as Person, Location, Organization, domain-specific Player, Team [14, 34, 46]. The system could jointly achieve F1-scores above 90% on IndicNER and WikiAnn multilanguage datasets with custom annotations, orders of magnitude better than previous Gujarati NER systems [2, 38, 39].

In this domain, an ontology was developed using Protégé, including classes for basic-domain entities, attributes, and relations such as playsFor, scored. Extracted entities were then mapped onto the ontology and stored in the form of RDF triples using RDFLib and Neo4j [8, 9, 23]. Ontology-driven IE provides semantic interpretation and means for structured querying via SPARQL, which is one of the major drawbacks of the purely neural approach [11, 19, 35].

Finally, a question-answering module was implemented to demonstrate the usefulness of this knowledge base. Gujarati queries like “વિરાટ કોહલીએ કેટલા રન બનાવ્યા?” (“How many runs did Virat Kohli score?”) would be parsed into entity-relation pairs, and SPARQL queries would be run over the RDF store to fetch exact answers [23, 24]. This integration proves the practical applicability of their system for tasks such as knowledge retrieval, decision support, and semantic search. This way, they demonstrated the deep learning and ontology-driven approach, which complement each other in achieving higher accuracy while delivering

structured and interpretable knowledge representations for Gujarati text.

Evaluation and Results

The next section describes the evaluation of the proposed BiLSTM-CRF + IndicBERT + Ontology architecture for Gujarati information extraction. Experiments were designed to cover POS tagging, chunking, and NER, as well as ontology-oriented storage and question answering.

Dataset Overview

The evaluation primarily used four datasets. UD Gujarati Treebank [12, 13] gave about 6,000 sentences (~120,000 tokens) with basic syntactic annotations. IndicNER dataset [14] gave about 10,000 annotation sentences with Persons, Locations, and Organizations, while WikiAnn Gujarati subset [41] provided 8,500 sentences. To cover terms specific to the domain, about 5,000 cricket news articles were annotated manually for entities like Player, Team. Altogether, the entire corpus contained approximately 300,000 tokens and 50,000 annotated entities, divided into 80% training, 10% validation, and 10% testing.

Evaluation Metrics

Several methods were used for evaluation, these including variables of accuracy for POS tagging and chunking, whereas NER was evaluated with the variables of precision, recall, and F1 score. Combined system accuracy was also reported for the whole pipeline. Training loss and accuracy were monitored over 20 epochs to check convergence and prevent overfitting.

Experimental Setup

Experiments were carried out on a powerful GPU server having NVIDIA Tesla V100 (32GB RAM) and Intel Xeon 64-core CPU processing units. Models were coded using PyTorch and HuggingFace Transformer libraries. Training was performed for 20 epochs with a batch size of 32, using Adam as the optimizer, with a learning rate of 3e-5, and a dropout rate of 0.3.

Model Performance

In Table 2, the overall system performances for the individual tasks are summarized. Detached from POS tagging and chunking, results show that both obtain accuracy scores of 96.10%, while NER gets a macro-average F1 score of 92.33%. The integrated pipeline was able to reach an overall accuracy of 96.48%, quite beyond that reported by the older Gujarati NER systems [2, 38, 39] that normally gave F1 scores below 80%.

Table - II Overall Model Performance

Task	Accuracy	Precision	Recall	F1-score
POS Tagging	96.10%	96.30%	95.80%	96.05%
Chunking	96.10%	96.00%	95.90%	95.95%
NER (Macro Avg.)	92.33%	92.80%	91.90%	92.33%
Overall System	96.48%	96.60%	96.20%	96.40%

Furthering the analysis of entity-level performance, precision, recall, and F1-scores per entity type are tabulated in Table 3. Domain-specific entities like Player, Team, and Verdict received the highest F1-score of 93.0%, thus proving the system's flexibility in addressing special domains.

Table - III NER Performance by Entity Type

Entity Type	Precision	Recall	F1-score
Person	93.1%	92.0%	92.5%
Location	92.5%	91.2%	91.8%
Organization	91.8%	91.0%	91.4%
Player/Team	93.7%	92.4%	93.0%
Average	92.8%	91.9%	92.33%

Evaluation Coverage

Generalizing is tested by drawing data from different fields. The cricket dataset had 5000 articles (~25,000 sentences). Likewise, 10000 sentences were randomly subsampled from IndicNER and WikiAnn. An uneven entity distribution was present across the corpora, with Person leading (40%), followed by Location (30%), Organization (20%), and domain-specific (10%).

Qualitative Results

Thus, the framework proves to set a strong ontologydriven representation. For example, a sentence was processed: "વિરાટ કોહલીએ 100 રન બનાવ્યા" (Virat Kohli scored 100 runs).

NER Output: Virat Kohli → PERSON, 100 run → SCORE

RDF Triple: <Virat_Kohli> <scored> <100_runs>

SPARQL Query: SELECT ?score WHERE {Virat_Kohli :scored ?score }

Answer: 100

This example demonstrates how unstructured Gujarati text can be converted into structured query-able knowledge, thereby aiding various downstream applications like question answering, decision support, and

semantic retrieval.

Comparative and Error Analysis

The proposed BiLSTM-CRF + IndicBERT + Ontology framework outperformed generic baselines such as CRF, HMM, and SVM, while also outperforming transformer only models, reporting an overall accuracy of 96.48%, whereas the earlier systems were in the range of 85– 90% [2, 7, 31, 38]. The integration of IndicBERT embeddings improved the contextual understanding, while ontology-based RDF triples facilitated semantic queries, thus bestowing the system with superiority in both accuracy and interpretability. However, some errors were noticed in ambiguous instances, such as “Gujarat High Court” being incorrectly labelled as LOCATION instead of ORGANIZATION, in code-mixed Gujarati-English text, and in infrequent domain-specific entities, such as less common player names or judicial terms. Instances requiring cultural or contextual knowledge also posed problems. These limitations set forth a case for a larger corpus, better-wrangling of code-mixed data, and incorporation of external knowledge sources in future.

Comparison With Existing Approaches

In order to show the effectiveness of the BiLSTM-CRF + IndicBERT + Ontology setup, a comparison was drawn between it and traditional ML baselines as well as the latest transformer models. Existing methods for Gujarati NLP have mostly involved CRF or rule-based systems with, additionally, recent works exploring multilingual pretrained transformers. However, most of them do not have semantic interpretability and domain-specific adaptability, an issue our framework addresses by means of ontology integration.

Existing Approaches

Table 4 gives a, kind of comparative look at different Gujarati Named Entity Recognition approaches, including the framework we propose. In earlier work, CRF based or rule-based pipelines achieved a decentish result, yet they also ran into trouble like limited generalization, and a narrow ability to adapt across domains. Later on, transformer-based solutions, for example IndicBERT, did help a lot with recognition accuracy because it learns multilingual context, but there’s a catch they typically need heavy compute, and their semantics are not easy to interpret.

In contrast, the proposed hybrid setup, which blends BiLSTM-CRF, IndicBERT, and ontology driven enhancement, delivers noticeably stronger outcomes, reporting 96.48% overall accuracy and 92.33% F1-score. Also, compared to other Gujarati NER systems, our approach shows better domain adaptability, and it feels more interpretable in terms of the underlying meaning.

Table IV Summary of Existing Approaches

Study / Model	Methodology	Dataset	Accuracy / F1	Strengths	Limitations
Shah & Bhadka [2, 38]	CRF + Rulebased	Gujarati News	~78% F1	Simple baseline	Low accuracy, poor generalization
Garg [1, 39]	CRF-based NER	Gujarati corpus	~80% F1	Early Gujarati NER	Limited entities, domain restricted
Vijayvergia et al. [21, 33]	Survey + Baselines	Indian Languages	~82% F1	Comparative insights	Few Gujarati specific benchmarks
IndicBERT [17]	Transformer (mBERT finetune)	IndicNER	~89% F1	Cross-lingual generalization	High compute cost, no ontology support
IndicNER [14], WikiAnn [41]	Multilingual dataset	News/Wiki	~90% F1	Broad coverage	Focused on basic entity types
Proposed Work (2026)	BiLSTM-CRF + IndicBERT + Ontology	UD_GujTB, IndicNER, WikiAnn, Cricket datasets	96.48% overall / 92.33% F1 (NER)	High accuracy, ontology driven interpretability, domain adaptability	Requires large annotated corpora and domain specific finetuning

Advantages of Proposed Work

The system makes improvements in many critical areas in comparison with previous systems. Rather than consider the CRF or rule-based approach [2, 38, 39], the framework, with BiLSTM feeding on the IndicBERT embeddings, considers sequential dependencies and contextual semantics for better tagging accuracy. In

contrast with transformer-only baselines [17, 18, 40], the model adds one ontology-driven application layer, whereby the entities and relations extracted can be represented as RDF triples and queried via SPARQL [8, 9, 23], thus rendering the system more amenable to semantic interpretation and adoption in real-world environments. Further, adaptation to domains was exhibited through performance on cricket and legal corpora with an overall accuracy of 96.48% and an NER F1-score of 92.33%, which is considerably higher than those of preexisting Gujarati systems that generally recorded less than 85% F1-score.

Quantitative Comparison

Table 5 gives this numerical sort of comparison between the proposed framework and the existing baseline methods, for Gujarati NLP tasks. The traditional CRF and those rule-based models basically ended up with smaller accuracy values and lower F1-scores, mainly because they struggled with contextual understanding, plus their semantic representations were not that strong. On the other hand, Transformer based approaches like IndicBERT and IndicBART show improved overall results, mostly because deep contextual learning really helps, especially for NER and POS tagging. Still, these approaches tend to come with rather limited explainability and semantic interpretation, so it is hard to reason about what exactly is happening. In contrast, the proposed ontology-driven hybrid framework beats every baseline, reaching 96.48% accuracy, 92.33% F1-score for NER, and 96.10% for POS/chunking, and it also adds better interpretability via RDF based ontology integration.

Table V Quantitative Comparison of Proposed Work vs. Baselines

Approach	Accuracy	F1-score (NER)	POS/Chunking	Explainability
CRF + Rulebased [2, 38]	~82%	~78%	~80%	Low
IndicBERT fine-tune [17]	~90%	~89%	~91%	Limited
Transformeronly (IndicBART) [18]	~91%	~90%	~92%	Limited
Proposed Work (2026)	96.48%	92.33%	96.10%	High (Ontology-driven RDF)

Discussion

As can be seen from Tables 3 and 4, in terms of accuracy and interpretability, the system appears to outperform any previous attempts at Gujarati NLP. Transformer-only approaches, with a F1 score of the third grade (~89-90%), do not provide semantic structuring and thereby hamper any downstream activities such as querying knowledge. Simultaneously, our ontology-driven approach gives the highest scores for POS, chunking, and NER and enables real-world applications in question answering and decision support.

Conclusion and Future Work

This paper proposed a hybrid deep-learning framework that combined BiLSTM-CRF models augmented with IndicBERT embeddings with ontology-driven representation to perform information extraction in the Gujarati language. The system tackled key tasks, such as part-of-speech tagging, chunking, and named entity recognition, extending those results into ontology-based RDF triples for semantic storage and SPARQL-based querying. Experimental evaluations on multiple datasets (categorized as UD Gujarati Treebank, IndicNER, WikiAnn, and custom domain corpora) show that our approach yields state-of-the-art performance with an overall accuracy of 96.48%, POS and chunking at 96.10%, and NER F1-score at 92.33%. The results vastly improve the earlier rule and CRF-only baselines [2, 38, 39] and, in addition, offer another level of interpretability and practicality through ontology integration.

Still, some hurdles remain despite all promising results. Such types of errors include code-mixed Gujarati-English text, rare domain-specific entities, or instances requiring contextual or cultural interpretations. There is also the issue of greater reliance on manually annotated data sets, which bottlenecked low-resource languages. Moreover, transformer embeddings are very expensive to calculate, thus limiting their scalability into lightweight deployments.

Future research would thus emphasize three principal directions. First, additional annotated corpora will need to be created for other domains such as finance, healthcare, and social media to increase generalization ability. Second, multimodal resources, including images, speech, and tabular metadata, should be incorporated to grasp the context richer and to apply it in a wider array of applications. Third, supplementing these efforts with better approaches to code-mix, low-resource scenarios, cross-lingual transfer learning, and multilingual fine-tuning will boost robustness. Lastly, integrating explainable AI methods with the ontology-driven framework might cement deeper interpretability, thereby establishing trust and transparency in decision support systems.

To conclude, thus, the formulation endorses that blending deep learning with ontology-driven representation is one of the best solutions for Gujarati information extraction in terms of great scalability, accuracy, and semantic interpretability. Other than Gujarati, this can be used as a blueprint for domain-adaptive NLP

systems for other low-resource Indic languages so that the common goal of inclusive and multilingual language technologies can be pursued further.

REFERENCES

1. Garg V, Saraf N, Majumder P. Named entity recognition for gujarati: a crf based approach In Mining Intelligence and Knowledge Exploration: First International Conference, MIKE 2013, Tamil Nadu, India, December 18-20, 2013. Proceedings 2013 Dec 18 (pp. 761-768). Cham: Springer International Publishing.
2. Shah DN, Bhadka HB. Named entity recognition from Gujarati text using rule-based approach. In International Conference on Intelligent Systems Design and Applications 2017 Dec 14 (pp. 797-805). Cham: Springer International Publishing.
3. Tailor C, Patel B. Hybrid pos tagger for gujarati text. In International Conference on Soft Computing and its Engineering Applications 2020 Dec 11 (pp. 134-144). Singapore: Springer Singapore.
4. Desai NP, Dabhi VK. Resources and components for Gujarati NLP systems: a survey. Artificial Intelligence Review. 2022 Oct;55(7):1-9.
5. Dave NR, Mehta MA, Kotecha K. Effective Stemmers Using Trie Data Structure for Enhanced Processing of Gujarati Text. International Journal of Computational Intelligence Systems. 2024 Dec 19;17(1):309.
6. Das BR, Patnaik S. A novel approach for Odia part of speech tagging using artificial neural network. In Proceedings of the international conference on frontiers of intelligent computing: Theory and applications (FICTA) 2013 2014 (pp. 147-154). Cham: Springer International Publishing.
7. Warjri S, Pakray P, Lyngdoh SA, Maji AK. Part-of-speech (POS) tagging using conditional random field (CRF) model for Khasi corpora. International Journal of Speech Technology. 2021 Dec;24(4):853-64.
8. Mo Q, Chen YH. Ontology-based web information extraction. In Communications and Information Processing: International Conference, ICCIP 2012 Aveiro, Portugal, March 7-11, 2012 Revised Selected Papers, Part I 2012 (pp. 118-126). Berlin, Heidelberg: Springer Berlin Heidelberg.
9. Karkaletsis V, Fragkou P, Petasis G, Iosif E. Ontology based information extraction from text. In Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap 2011 May 6 (pp. 89-109). Berlin, Heidelberg: Springer Berlin Heidelberg.
10. Wimalasuriya DC, Dou D. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science. 2010 Jun;36(3):306-23.
11. Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I. Information extraction meets the semantic web: a survey. Semantic Web. 2020 Feb 5;11(2):255-335.
12. Jobanputra M, Mehta M, Çöltekin Ç. A Universal Dependencies Treebank for Gujarati. In Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024 2024 May (pp. 56-62).
13. Universal Dependencies, "UD_GujaratiGujTB Treebank," 2024. [Online]. Available: <https://universaldependencies.org>
14. Mhaske A, Kedia H, Doddapaneni S, Khapra MM, Kumar P, Murthy R, Kunchukuttan A. Naamapadam: A large-scale named entity annotated data for Indic languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2023 Jul (pp. 10441-10456).
15. Kunchukuttan A, Kakwani D, Golla S, Bhattacharyya A, Khapra MM, Kumar P. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. arXiv preprint arXiv:2005.00085. 2020 Apr 30.
16. Kunchukuttan A, Kakwani D, Golla S, NC G, Bhattacharyya A, Khapra MM, et al. AI4Bharat-IndicNLP Corpus [Internet]. GitHub; 2020. Available from: https://github.com/AI4Bharat/indicnlp_corpus
17. AI4Bharat. IndicBERT: Pretrained multilingual language model for Indic languages [Internet]. GitHub; 2020. Available from: <https://github.com/AI4Bharat/Indic-BERT-v1>.
18. Dabre R, Shrotriya H, Kunchukuttan A, Puduppully R, Khapra MM, Kumar P. IndicBART: A pre-trained model for indic natural language generation. In Findings of the Association for Computational Linguistics: ACL 2022 2022 May (pp. 1849-1863).
19. Bhogayata CK. A Machine Learning-Based Readability Model for Gujarati Texts. ACM Transactions on Asian and Low-Resource Language Information Processing. 2024 Feb 8;23(2):1-32.
20. Deode S, Gadre J, Kajale A, Joshi A, Joshi RB. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation 2023 Dec (pp. 154-163).
21. Vijayvergia R, Nathani B, Joshi N, Jain R. A survey on various approaches used in named entity recognition for indian languages. In Proceedings of the 4th International Conference on Information Management & Machine Intelligence 2022 Dec 23 (pp. 1-5).
22. Bhattacharjee K, Mehta S, Kumar A, Mehta R, Pandya D, Chaudhari P, Verma D. Named entity

- recognition: A survey for indian languages. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT) 2019 Jul 5 (Vol. 1, pp. 217-220). IEEE.
23. Tao J, Deokar AV, El-Gayar OF. An ontology-based information extraction (OBIE) framework for analyzing initial public offering (IPO) prospectus. In 2014 47th Hawaii International Conference on System Sciences 2014 Jan 6 (pp. 769-778). IEEE.
24. Shim M, Choi H, Koo H, Um K, Lee KH, Lee S. OmEGa (Ω): Ontology-based information extraction framework for constructing task-centric knowledge graph from manufacturing documents with large language model. *Advanced Engineering Informatics*. 2025 Mar 1;64:103001.
25. Sheshadri SK, Gupta D, Costa-Jussà MR. A voyage on neural machine translation for indic languages. *Procedia Computer Science*. 2023 Jan 1;218:2694-712.
26. Javed T, Nawale J, George E, Joshi S, Bhogale K, Mehendale D, Sethi I, Ananthanarayanan A, Faquih H, Palit P, Ravishankar S. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024* 2024 Aug (pp. 10740-10782).
27. De Marneffe MC, Manning CD, Nivre J, Zeman D. Universal dependencies. *Computational linguistics*. 2021 Jun;47(2):255-308.
28. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*. 2023 Jan;82(3):3713-44.
29. Panchal BY, Shah A. A survey on Gujarati NLP research work. *Aibi Rev Investig Adm Ing*. 2025;13(1):234-249. doi:10.15649/2346030X.4445.
30. Bhatt PM, Ganatra A. POS-HOML: POS tagging technique for Gujarati language using hybrid optimal and machine learning approaches. *Int J Eng Trends Technol*. 2021;69(11):256-261.
31. Baxi J, Soni O, Bhatt B. Part of Speech and Morph Category Prediction for Gujarati. *Int J Intell Syst Appl Eng [Internet]*. 2024 Mar. 24 [cited 2026 Jun. 20];12(3):586-99. Available from: <https://ijisae.org/index.php/IJISAE/article/view/5290>
32. Vijayvergia R, Nathani B, Joshi N, Jain R. A survey on various approaches used in named entity recognition for indian languages. In *Proceedings of the 4th International Conference on Information Management & Machine Intelligence 2022* Dec 23 (pp. 1-5).
33. Prabhakar A, Majumder GS, Anand A. CL-NERIL: A cross-lingual model for NER in indian languages (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence 2022* Jun 28 (Vol. 36, No. 11, pp. 13031-13032).
34. Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I. Information extraction meets the semantic web: a survey. *Semantic Web*. 2020 Feb 5;11(2):255-335.
35. Warjri S, Pakray P, Lyngdoh SA, Maji AK. Part-of-speech (POS) tagging using conditional random field (CRF) model for Khasi corpora. *International Journal of Speech Technology*. 2021 Dec;24(4):853-64.
36. Das BR, Patnaik S. A novel approach for Odia part of speech tagging using artificial neural network. In *Proceedings of the international conference on frontiers of intelligent computing: Theory and applications (FICTA) 2013* 2014 (pp. 147-154). Cham: Springer International Publishing.
37. Shah DN, Bhadka HB. Named entity recognition from Gujarati text using rule-based approach. In *International Conference on Intelligent Systems Design and Applications 2017* Dec 14 (pp. 797-805). Cham: Springer International Garg V, Saraf N, Majumder P. Named entity recognition for gujarati: a crf based approach. In *Mining Intelligence and Knowledge Exploration: First International Conference, MIKE 2013, Tamil Nadu, India, December 18-20, 2013. Proceedings 2013* Dec 18 (pp. 761-768). Cham: Springer International Publishing.
38. Dabre R, Shrotriya H, Kunchukuttan A, Puduppully R, Khapra MM, Kumar P. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022* 2022 May (pp. 1849-1863).
39. AI4Bharat. AI4Bharat GitHub organization [Internet]. Available from: <https://github.com/AI4Bharat>
40. Jobanputra M, Mehta M, Çöltekin Ç. A Universal Dependencies Treebank for Gujarati. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024* 2024 May (pp. 56-62).
41. Desai NP, Dabhi VK. Resources and components for Gujarati NLP systems: a survey. *Artificial Intelligence Review*. 2022 Oct;55(7):1-9.
42. Khanam MH, Khudhus MA, Babu MP. Named Entity Recognition using Machine learning techniques for Telugu language. In 2016 7th IEEE international conference on software engineering and service science (ICSESS) 2016 Aug 26 (pp. 940-944). IEEE.
43. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th international conference on computational linguistics 2018* Aug (pp. 2145-2158).
44. Bhattacharjee K, Mehta S, Kumar A, Mehta R, Pandya D, Chaudhari P, Verma D. Named entity

- recognition: A survey for indian languages. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICT) 2019 Jul 5 (Vol. 1, pp. 217-220). IEEE.
45. Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I. Information extraction meets the semantic web: a survey. *Semantic Web*. 2020 Feb 5;11(2):255-335.
 46. Deode S, Gadre J, Kajale A, Joshi A, Joshi RB. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation* 2023 Dec (pp. 154-163).
 47. Mhaske A, Kedia H, Doddapaneni S, Khapra MM, Kumar P, Murthy R, Kunchukuttan A. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2023 Jul (pp. 10441-10456).