RESEARCH ARTICLE OPEN ACCESS

Temporal Information Extraction in Gujarati Text using a Rule-Based Approach

Parul Patel
(M.Sc(IT) Programme, VNSGU, Surat
India

ABSTRACT
Temporal information extraction (TIE) plays a critical role in natural language processing (NLP) applications such as question answering, information retrieval, and event timeline generation. While significant research exists for English and other resource-rich languages, temporal analysis for low-resource languages like Gujarati remains unexplored. This paper presents a rule-based approach for temporal information extraction in Gujarati text, focusing on identifying and normalizing expressions related to time (dates, days, duration, temporal adverbs). Our system leverages handcrafted linguistic rules based on Gujarati grammar, suffix patterns, and temporal keywords. Preliminary evaluations on a curated dataset of Gujarati news articles and stories demonstrate the effectiveness of rule-based methods in handling morphological rich language features, achieving promising accuracy for explicit temporal expressions.

Keywords :— Temporal Information Extraction, Gujarati NLP, Rule-Based Approach, Date Normalization, Low-Resource Language

I. INTRODUCTION

Temporal information refers to expressions in text that denote time, including dates, times, durations, and temporal signals etc.

Type	Example	English Meaning
Date	"૧ જાન્યુઆરી ૨૦૨૫", "આવતા રવિવારે", "પાછલા મહિને"	1 Jan 2025, next Sunday, last month
Time	"સવારના ૬ વાગ્યે", "બપોરે", "રાત્રે ૯:૩૦", "મધરાત્રી"	6 AM, afternoon, 9:30 PM, midnight
Duration	"બે કલાક", "ત્રણ દિવસ", "એક સપ્તાહ", "ઘણાં વર્ષો સુધી"	2 hours, 3 days, 1 week, for many years
Frequency	"દરરોજ", "સપ્તાહમાં બે વાર", "મહિને એક વાર", "વારંવાર"	daily, twice a week, once a month, frequently
Relative	"આજે", "ગઈકાલે", "આવતી કાલે", "હમણાં", "થોડીવાર પછી"	today, yesterday, tomorrow, now, after a while
Cultural/Seasonal	"દિવાળી પછી", "ઉત્તરાયણના દિવસે", "પાડવામાં", "ચોમાસા દરમિયાન"	after Diwali, on Uttarayan, on Gujarati New Year, during monsoon

Extracting such information is vital for applications like temporal question answering, event ordering, and historical data analysis. For high-resource languages, machine learning and deep learning approaches dominate temporal information extraction. However, Gujarati—a morphologically rich Indo-Aryan language—suffers from a lack of annotated corpora and computational resources, making purely data-driven methods less feasible.

This research explores a rule-based approach tailored for Gujarati text. We design linguistic rules to detect explicit temporal expressions, handle morphological variations, and normalize temporal information into a structured format (e.g., ISO 8601).

II. LITERATURE REVIEW

Rule-based methods, though less scalable than machine learning, remain effective for low-resource languages where linguistic knowledge can compensate for data scarcity.The ParsTime system is based on rules, which enables it to extract and standardize Persian time-related expressions as per TIMEX3 annotation guidelines. Experimental outcomes demonstrate that ParsTime can identify temporal expressions in Persian texts with an F1-score of 0.89[1].Research in [2] focuses on temporal expression (TE) extraction and

normalization challenges in Chinese clinical notes using a rule-based system. The system, divided into direct, indirect, and uncertain TEs, achieved an F-score of 93.40% on TE extraction and 92.58% on TE normalization. TimeBankPT[3] is a corpus developed for Portuguese language annotated in TimeML format. INDTime[4] is a temporal tagger developed for English focusing on festivals, and events in Indian context. Research work in [5] presents rule based and machine learning based approaches for identifying and classifying temporal entities in Hindi. CRF-based classifier is used trained on human tagged data, achieving a strict F1-measure of 0.78. The ILTIMEX2012 corpus, a gold standard dataset, is used for temporal tagging.KtimeML[6] provides Specification of Temporal and Event Expressions in Korean Text. Based on literature work, it is observed temporal information extraction has been widely studied in English, especially through frameworks like **TimeML [7]** and **HeidelTime[8]**. In Indian languages, limited work exists:

- Hindi and Bengali have seen preliminary systems for temporal tagging using hybrid rule-based approaches.

Gujarati NLP research has primarily focused on tokenization, morphological analysis, and part-of-speech tagging, with minimal emphasis on temporal extraction.

III. RESEARCH METHODOLOGY

Due to morphological richness of Gujarati language, it is important to understand the linguistic characterstics of it. Temporal expression may appear at any position in the sentence. Some temporal expression require context to interpret like "*કાલે*" which can mean “yesterday” or “tomorrow,” depending on context. To resolve ambiguity in such expression, understanding of entire sentence becomes necessary. In this context, we have first manually analyzed 500 gujarati news articles, blogs, and short stories. After analysis, sentences focusing on temporal expressions were extracted, and a collection of 1,500 such sentences was developed. Sentences containing explicit time expressions were manually annotated.

Figure 1 depicts flow diagram of temporal information extraction pipeline. Based on annotated corpus, regular expression-based on linguistic rules are designed to detect temporal categories:

(1) Regular expressions are designed to extract explicit temporal expressions, focusing on day, , month, year and date etc. For e.g. Pattern: `[0-9]{1,2}[-/.][0-9]{1,2}[-/.][0-9]{2,4}` (e.g., *12-08-2015*). Gujarati Month like: જાન્યુઆરી, ફેબ્રુઆરી, માર્ચ,

(2) Regular expressions for day like (keywords: સોમવાર, મંગળવાર, બુધવાર...) *with* Morphological handling for phrases (સોમવારે, મંગળવારે...) are designed.

(3) Regular expression focusing on relative expressions like (કાલે (yesterday/tomorrow), ગઈકાલે (yesterday), આજે (today), હમણાં (now), ગયા મહિને (last month)) are designed.

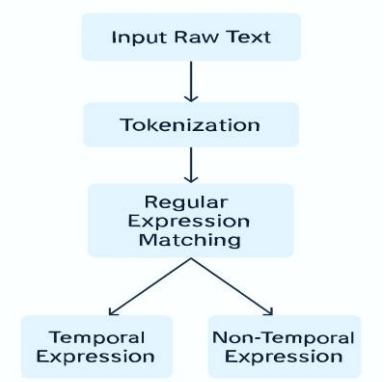


Figure 1. Temporal Information Extraction Pipeline

(4) Regular expression focusing on duration are designed (Numeric + Unit: ૩ દિવસ, બે અઠવાડિયા, ૫ વર્ષ.)

All explicit temporal expressions are converted into date format (DD-MM-YYYY). All relative expressions are mapped to relative offsets from document creation time (DCT). All regular expressions are designed using java regex

IV. EVALUATION & RESULTS

For evaluation, the manually annotated dataset described in Section II was used. The dataset contains 1000 test sentences. Evaluation metrics like precision, recall and f-measure are used.

Precision (P) measures correctness of the system’s results.

$$P = \frac{Correctlyextractedtemporalexpressions(TP)}{Allextractedexpressions(TP+FP)} \quad (1)$$

TP- True Positive
FP- False Positive
FN- False Negative

Recall measures the coverage of the system.

$$R = \frac{Correctlyextractedtemporalexpressions(TP)}{Allactualtemporalexpressions(TP+FN)} \quad (2)$$

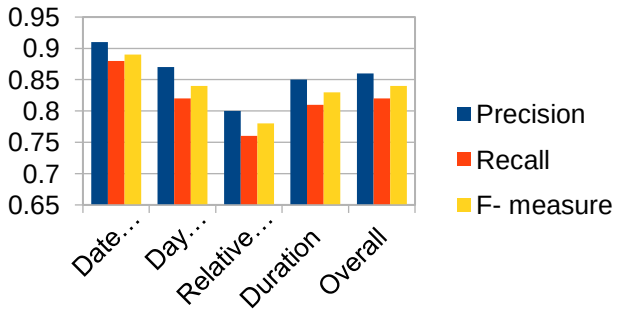
F measure is a mean of precision and recall.

$$F = \frac{2XPrecisionXRecall}{Precision+Recall} \quad (3)$$

Table 2. Result

Category	Precision	Recall	F- measure
Date Expressions	0.91	0.88	0.89
Day Expressions	0.87	0.82	0.84
Relative Expressions	0.80	0.76	0.78
Duration	0.85	0.81	0.83
Overall	**0.86**	**0.82**	0.84

Figure 2. Comparison of result for extracted temporal expression



The rule-based approach proves effective for explicit temporal expressions in Gujarati. However, challenges remain for:

- Ambiguity resolution (કાલે).
- Implicit temporal references (પાછલા વર્ષે દિવાળી "last year's Diwali").
- Domain adaptation (news vs. conversational text).

V. CONCLUSION & FUTURE WORK

This paper presents a rule-based framework for temporal information extraction in Gujarati text. Despite resource limitations, linguistic rules achieve strong performance on explicit temporal expressions. This work contributes towards foundational NLP tools for Gujarati and sets the stage for hybrid or neural approaches as annotated datasets become available. A hybrid system integrating rule-based extraction with machine learning models for disambiguation could further improve accuracy.

REFERENCES

[1] Mansouri, Behrooz & Zahedi, Mohammad & Campos, Ricardo & Farhoodi, Mojgan & Rahgozar, Maseud. (2018). ParsTime: Rule-Based Extraction and Normalization of Persian Temporal Expressions. 10.1007/978-3-319-76941-7_67.

[2] Liu, Z., Tang, B., Wang, X., Chen, Q., Li, H., Bu, J., Jiang, J., Deng, Q., & Zhu, S. (2017). CMedTEX: A Rule-based Temporal Expression Extraction and Normalization System for ChineseClinical Notes. AMIAAnnual Symposium proceedings. AMIA Symposium, 2016, 818–826.

[3] Costa, Francisco and A. Branco. "TimeBankPT: A TimeML Annotated Corpus of Portuguese." LREC (2012).

[4] Patel P. & Patel, S. (2016). INDTime: Temporal Tagger––First Step Toward Temporal Information Retrieval. 10.1007/978-981-10-0129-1_21.

[5] Ramrakhiyani N., Majumder P. (2013) Temporal Expression Recognition in Hindi. In: Prasath R., Kathirvalavakumar T. (eds) Mining Intelligence and Knowledge Exploration. Lecture Notes in Computer Science, vol 8284. Springer, Cham. https://doi.org/10.1007/978-3-319-03844-5_72

[6] S. Im, H. You, H. Jang, S. Nam, H. Shin, "KTimeML: specification of temporal and event expressions in Korean text," in Proceedings of the 7th Workshop on Asian Language Resources, Singapore, 2009; pp. 115-122

[7] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[8] Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.