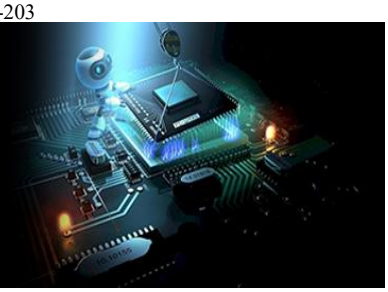
International Journal of Engineering in Computer Science



**E-ISSN:** 2663-3590
**P-ISSN:** 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2025; 7(1): 200-203
Received: 02-03-2025
Accepted: 06-04-2025

**Anupama Sharma**
Department of computer science and engineering, Punjabi University, Patiala, Punjab, India

**Dr. Madan Lal**
Department of computer science and engineering, Punjabi University, Patiala, Punjab, India

**Corresponding Author:**
**Anupama Sharma**
Department of computer science and engineering, Punjabi University, Patiala, Punjab, India

# Temporal information retrieval in Sanskrit using Rule-based approach

**Anupama Sharma and Madan Lal**

**DOI:** https://www.doi.org/10.33545/26633582.2025.v7.i1c.179

**Abstract**
Identification of temporal expressions is a significant task in NLP (Natural language processing), especially when it comes to comprehending time-related expressions in texts. This study provides a rule-based approach to extract temporal expressions in Sanskrit language. The accuracy of the system is accessed on a selected dataset by using regular expressions (regex) to correlate preset temporal keywords along with their morphological variants.

**Keywords:** Temporal information retrieval, Sanskrit language processing, Time expression extraction, Rule based approach

## Introduction

Retrieval of temporal information is receiving a lot of attention recently. By taking into account temporal information in text it aims to enhance the efficacy of information retrieval techniques. Words and phrases with time significance are referred to as temporal expressions [1]. This study utilizes regex to construct a rule-based system for temporal expression detection. For the languages like Sanskrit with sparse annotated data, rule-based algorithms work well. It is challenging to retrieve temporal expressions in Sanskrit due to morphological complexity and contextual usage as shown in figure 1.

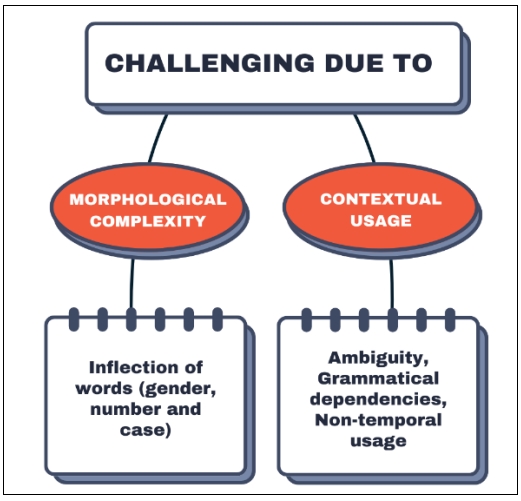



**Fig 1:** Challenges of temporal expressions retrieval in Sanskrit.

## Related work

In order to recognize and standardize temporal phrases in clinical texts published in Spanish, Najafabadipour *et al* [2] developed a temporal tagger. Various tasks involved in the tagger are extracting, filtering, resolving and normalising various temporal expressions. Vichayakitti [3] reported rule-based method to identify temporal events in Thai text news. The problem is broken down into three categories: algorithm implementation, part-of-speech tagging, and word segmentation.

The text is annotated with five different sorts of tags: date, number, sentence end, time-related terminology, and pertinent events. Martinez-Barco [4] described phrase production, question answering, event analysis and translation as various applications of temporal expression identification. To find these phrases and annotate the corpus of words, timex with a time noun specified by a time adverb are analysed. Mani [5] These outlined techniques for extracting time expressions from various languages and presents a set of rules for annotating them with a canonical form of the time they relate to. Mazur, Dale [6] discussed DANTE system to recognize as well as normalize temporal expressions of English language by using Timex2 standards. Different tasks involved are tokenizing, sentence splitting, POS tagging, recognising named entity, identifying and interpreting temporal expression. Ahmed *et al* [9] developed a co-trained model for retrieving news articles with a focus on time in order to improve study in semi-supervised learning. The dataset is mapped by examining two primary aspects: the evolution of news item emphasis over time, and a semi-supervised learning technique that learns simpler patterns by using relevant context. The created model achieved 89% output with semi supervised technique as well as lexicon extension.

**Table 1:** Temporal expression systems developed for different languages, corresponding datasets and results

| Author | Language | Dataset/Subject system | Results |
|---|---|---|---|
| Najafabadipour [2] | Spanish | Medical domain (electronic health records of individuals with lung cancer) | F1 score 0.93 |
| Vichayakitti [3] | Thai | Thai Newspapers | Accuracy 68% |
| Martinez-Barco [4] | Spanish | Wikipedia in Spanish, News articles | N/A |
| Mani *et al* [5] | English, Spanish | Corpus containing 32000 words of telephone dialog and for Spanish, Enthusiast corpus | F-score is 96.2% |
| Mazur, Dale [6] | English | ACE 2005, ACE 2007 corpus | N/A |
| Negri and Marseglia [7] | English | N/A | Precision is 97.6%, Recall is 80% |

## Methodology

Figure 2 shows the steps involved in identifying the temporal expressions.

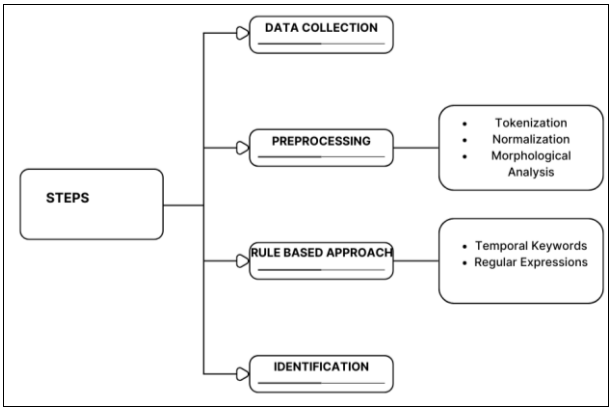



**Fig 2:** Steps for recognising temporal expressions.

### Data collection

Data is collected from various sources to cover various temporal expressions in Sanskrit**.**

### Preprocessing

It requires three steps.
Tokenization: Divide the text into discrete tokens or words.
Normalization: Cleaning as well as standardizing the text.
Morphological Analysis: Dissect words based on their suffixes and root forms.

### Rule based approach

### Temporal keywords

A list of temporal keywords is compiled which covers small, intermediate and large temporal units. For example, परमाणु, क्षण, मुहूर्त, दिन, रात्रि, सप्ताह, मास. Also, adverbs (of time) are listed for example "अद्य", "ह्यः", "श्वः".

**Regular expressions:** Following the listing of keywords, we employ regular expressions to generate patterns that correspond to these terms and their variants in phrases.

- Suffixes are appended to each term to accommodate frequent morphological forms. For example, depending on the situation or context, "दिनरात्र" could end in "ौ," "े," or "स्य," while "वर्ष" may appear as "वर्षे" or "वर्षस्य."
- To match combinations of digits and temporal units, regex patterns were created.
- The OR operator (|) in regular expression is then used to combine these keywords and their variants, resulting in a single pattern which could match either of the specified temporal phrases in the sentence.

### Identifying temporal expressions

To look for matches to the regex pattern in text, a function called identify temporal expression is developed. When given the text, the function looks for any instances of the temporal keywords using the regex pattern and produces a list of results. The text's temporal expressions are represented by any time-related terms that are discovered.

### Dataset

A dataset of 100 phrases was gathered. A ground truth for assessment was provided by the manual annotation of each phrase for temporal expressions.

### Implementation

Regular expressions in Python are used in the implementation to match temporal expressions. Important characteristics include regular expression matching and evaluation metrics.

### Evaluation metric

Following evaluation metric are used to evaluate the system.

### Precision

Calculates the number of correctly detected temporal expressions.

**Recall**
Quantifies the number of real temporal expressions that the algorithm was able to identify.

**True positive**
When the algorithm accurately detects a temporal expression, it is called a True Positive.
The ground truth and system output are same.

**False positive**
When expression is incorrectly recognized by the system as a temporal expression when it is not, this is known as a false positive.
**Ground Truth:** There isn't any time expression.
**System Output:** A temporal expression is misidentified by the system.

**False Negative (FN)**
When a temporal expression that is truly present is not detected by the system, this is known as a False Negative.

**True Negative (TN)**
When a statement that actually has no temporal expression is accurately identified by the system as having no temporal expression, this is known as a True Negative.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**Table 2:** Evaluation metrics

| Case | Definition | Correctness |
|---|---|---|
| True positive | Accurately recognizes a time expression | ☑ |
| False positive | Incorrectly recognizes a non-temporal expression | ✗ |
| False negative | Fails to recognize a true temporal expression | ✗ |
| True negative | Appropriately neglects a statement that lacks temporal expressions | ☑ |

**Results and Discussion**
True positive= 47, False positive=25, False negative=22, Precision ≅ 0.65, Recall ≅ 0.68, F-score ≅66.

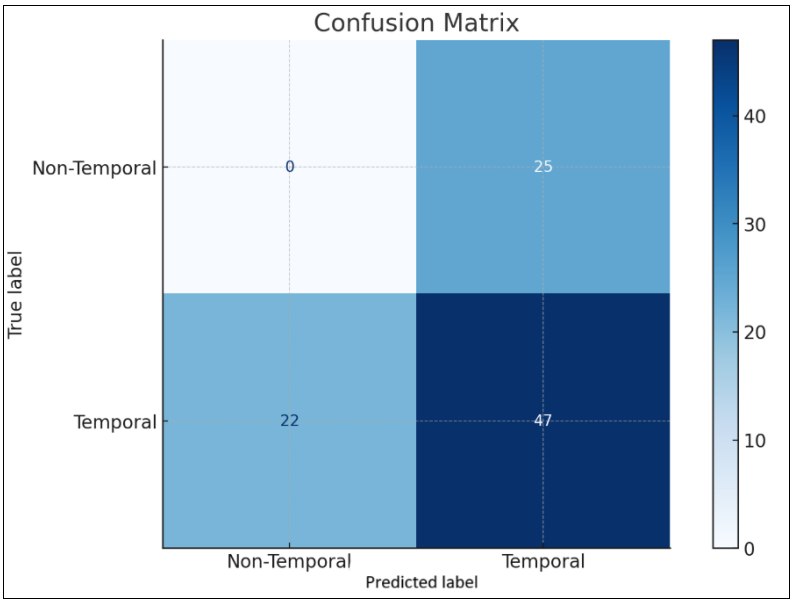



**Fig 3:** Confusion matrix

**Conclusion**
This study demonstrates rule-based approach for identifying temporal expression for Sanskrit by utilizing the rich morphological structure and linguistic aspects of the language. Our approach captured temporal expressions inside the language's rich and complex syntactic structure by combining morphological analysis, linguistic expertise, and hand-crafted rules. The method offers a repeatable foundation that may be modified for additional languages that are morphologically rich, despite constraints like contextually dependent ambiguities and the versatility of rule sets. In order to improve generalizability, future research may use machine learning techniques and increase the dataset's coverage.

**Conflict of Interest:** The authors state that they have no financial, personal, authorship, or other conflicts of interest that might affect the research and findings reported in this publication.

**References**
1. Bansal R, Rani M, Kumar H, Kaushal S. Temporal information retrieval and its applications: A survey. In: Proceedings of Emerging Research in Computing, Information, Communication and Application. 2019;2:251-262.
2. Najafabadipour M, López Ú, Mena G, Bessis N, Martí R. Recognition of time expressions in Spanish electronic health records. In: Proceedings of IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). Córdoba, Spain; 2019. p. 69-74. DOI:10.1109/CBMS.2019.00025.

3. Vichayakitti T, Jaruskulchai C. Automatic temporal event recognition from Thai news. In: IEEE International Symposium on Communications and Information Technology (ISCIT). China; 2005. p. 938-942. DOI:10.1109/ISCIT.2005.1567021.
4. Martinez-Barco P, Saquete E, Muñoz R. A grammar-based system to solve temporal expressions in Spanish texts. In: Proceedings of International Conference for Natural Language Processing. Portugal; 2002. vol. 2389. https://doi.org/10.1007/3-540-45433-0_8
5. Mani I, Wilson G, Sundheim B, Ferro L. A multilingual approach to annotating and extracting temporal information. In: Proceedings of the Workshop on Temporal and Spatial Information Processing. Association for Computational Linguistics; 2001;13(12):1-7.
6. Mazur P, Dale R. The DANTE temporal expression tagger. In: Proceedings of Human Language Technology. Challenges of the Information Society: Third Language and Technology Conference. Poland; 2007. p. 245-257.
7. Negri M, Marseglia L. Recognition and normalization of time expressions: ITC-irst at TERN 2004.
8. Mazur P. Temporal expression tagger. In: Human Language Technology Challenges of the Information Society. Springer; 2009. p. 245-257.
9. Ahmed U, Lin JCW, Diaz VG. Automatically temporal labeled data generation using positional lexicon expansion for focus time estimation of news articles. ACM Trans Asian Low-Resour Lang Inf Process. 2024;23(5):1-20.