

Temporal Information Identification and Normalization from Multilingual Social Media

Anjana Patel¹, Vira Bagiya², Sandip Modha³

Assistant Professor^{1, 2, 3}

Department of Information Technology

GEC, Gandhinagar Gandhinagar, Gujarat, India¹, OM Engineering College Junagadh, Gujarat²

Kadi Sarva Vishwavidyalaya Gandhinagar, Gujarat, India³

Corresponding Authors' email id: *anjanapatel30@gmail.com*

Abstract

Social media generates massive amount of multilingual real time data. On Many occasion, Social media have proved that it can be fastest media to break event compare to any other traditional news media e.g. News channel, Internet news portal. In this work, the identification and the normalization of temporal information have been done on tweets. Each tweet can represent or describe event as opposed to previous work which rely of bursty keyword detection. In the Identification of temporal entity, three approach: Conditional Random Field, Support Vector Machine and Rule base approach used. In the Normalization of temporal entity, Rule base approach used. The rule base approach proceeds texts as inputs and from various rules, it identify temporal information. The conditional approach is trained using tagged data and then using generated model tagging on test dataset done. Morphological, Syntactic and gazetteers features are defined in Conditional random field. The use of chunking in Conditional Random Filed is not effect in identification of temporal entity. In support vector machine approach, training is completed on tagged dataset and then using generated model file temporal entity tagged. The result on Hindi temporal named entity identification terms of Precision 0.88, Recall 0.81, F1 - measure 0.85 for rule base approach; Precision 0.93, Recall 0.74, F1 - measure 0.83 for conditional random field; Precision 0.73, Recall 0.69, F1 - measure 0.71 for support vector machine

and its normalization in terms of Precision 0.85, Recall 0.80 and F1-measure 0.83. Rule base approach gives good result compare to other two approach for temporal entity identification. Temporal expression identification can be helpful for generation tweet calendar, Question/Answering over social media and event summarization.

Keywords: Hindi temporal, temporal expression identification

INTRODUCTION

Social media is useful for disseminate information about people, services, products, institutes, governments and organizations. Social media e.g. Facebook, Twitter become richest source of real time information due to advancement in ICT technologies. Smart phone revolution in India make social media an important communication medium in individual's life. Social media allow us to post or disseminate text/image/video related to any personal event/festival/or natural disaster. Some researcher describe that social media empower human to act as a social sensor [3]. People are using smart phone, tablets frequently for communicating on social media.

Many Popular celebrities choose twitter to break their personal event like marriage, promotion of music, etc. There are many occasion where twitter is the first media were new event breaks. People are more contented in their native language for

interactive on social media. Since, Twitter support 6 Indian languages (like Hindi/Gujarati/Bengali/Tamil/Marathi/Kannada), Social Media user has started tweeting in this Indian language. Though number of tweet in Indian language are very less but these are increasing exponentially. Each tweet indicate some events. Events has four attribute: Who, When, What and Where. Who designate named entity e.g person name, location, organization name. When specify temporal information e.g. period, date, time and set. Where specify location entity. What designate specifically event. Compare to European language, the information in Indian language are in small amount. Till currently numerous approaches have been established to achieve temporal expressions for European language, but for Indian language there is not that much sufficient work done [12]. Therefore, identification of temporal information over Indian languages occurs for Tweets. Temporal expression identification and

normalization is one of subtask of Information Retrieval. This identification of temporal expression is used in application like when kind question answering, event identification. I have chosen Hindi Tweet written in Devanagari script for our experiment. Hindi is 3rd most spoken language in the world [11].

The temporal entity e.g. interval or a time are called a temporal expression (Timex) [13]. For example, PM Modi a tweeted "Congratulations & best wishes to Ms. Mehbooba Mufti, Dr. Nirmal Singh & all those who took oath today" and in Hindi देखिए रुहानी सत्संग का सीधा प्रसारण, 25 जनवरी 2016, रात 08 बजे से सिर्फ सर्वधर्म संगम चैनल पर. The phrase today, 25 जनवरी 2016 and रात 08 बजे are called temporal expression.

Temporal expressions are three types [6]: fully qualified, deictic and anaphoric. Timex can be unambiguously refer to an accurate interval or point in the time domain, it is called as fully qualified. "रात 08 बजे", "20 जनवरी 1950", "2016 से प्रारंभ होकर शुक्रवार 01 अप्रैल, 2016" are examples of fully qualified. Timex which has binding with the time domain- TCT (Tweet Creation Time) are called as deictic expressions. "आज", "कल", "पिछली

23 जनवरी" are example of deictic temporal expressions. Anaphoric expressions are a deictic expressions for which the TCT varies according to the Timex previously mentioned in the sentences. "पिछले महीने", "अगले साल" are example of anaphoric expressions.

The main role of this work is the identification and normalization of temporal expression in Hindi. We have inspired by the work of author [7, 11, 12] which take Hindi article as input. I have extended his work on tweet and furthermore, I have done normalization part. This work is differing from popular research work carried out by [1] [4] [6] in following way (i) I have taken every single tweet as event candidates by exploring its temporal attributes. (ii) I have taken Hindi tweet to detect temporal expressions, till now no one has worked on Hindi tweet.

The paper is organized in following way: Section 2 discusses related work, section 3 discusses the problem formulation, section 4 discuss corpus for IETE FIRE 2016 track. Section 5 discuss our approach, section 6 discusses result on IETE FIRE 2016 corpus.

2 Related works

The idea of automatically annotating temporal expressions in texts appeared for

the first time in 1998 [3]. The temporal expression extraction problem has two subtask: identification phase and normalisation phase [6]. In the first subtask, detection of the right boundary of temporal expressions is done. In the 2nd subtask, the temporal meaning of each pre-identified expression is interpreted and represented using the TimeML format [13].

For identification, M Filannino et al. [6] used morphological features with Conditional Random Field (CRF) [5]. The identification performance ranges from 0.81 to 0.90 in terms of lenient F-1 measure and from 0.70 to 0.83 for strict matching in TempEval-3 [10]. A variety of features have been used such as morphological and dictionary-based.

For normalisation rule-based approaches used. UzZaman and Allen [10] developed TRIOS, which focussing on type and value attributes. The Normalisation performance range from 0.68 to 0.86 for value and 0.86 to 0.94 for type attribute [10].

In the TempEval-3 challenge, ManTIME [6] system is based on CRF and rule base approach for temporal information identification and normalization. More recently, HeidelTime [4] proposed rule based approach, SUTime [1] use a different approach: a deterministic rule-

based system built on top of the Stanford Core NLP pipeline.

3 Problem Formulations

The temporal entity e.g. interval or a time are called a temporal expression (Timex) [13]. A Temporal entity contains one or more pivot words. Quantifiers quantify or modify the pivot word. Direction words give a time direction to the pivot word [7].

15 जनवरी 1950, १० साल are the examples of the temporal expression. Temporal expression can be four types: Date, Time, Period and set. Date can be combination of day, month and year. Time can be in form of hh:mm:ss or hh:mm or (digit बजे) e.g. प्रातःकाल 11:00, सुबह 10 बजकर 18 मिनट. Period indicate any duration e.g. 9 साल और 8 महीने, ४८ घंटे, 18-24 जनवरी 2016. Set is time occur with some frequency e.g. हर महीने. Hindi temporal words e.g. names of days, date and time formats, name of months and name of seasons form pivot words [11]. The aim of the system is to fulfil following goals:

1) Temporal Expression Identification
The temporal expression, also called Timex, refers to any natural language phrase denoting a temporal entity such as an interval or a time point [13]. In

temporal identification task, aim is to find right boundary of temporal expressions. “Congratulations & best wishes to Ms. Mehbooba Mufti, Dr. Nirmal Singh & all those who took oath today”. “देखिए रुहानी सत्संग का सीधा प्रसारण, 25 जनवरी 2016, रात 08 बजे से सिर्फ सर्वधर्म संगम चैनल पर” . The phrase today and 25 जनवरी 2016, रात 08 बजे is called temporal expression.

TIMEX is stand for Time Expression. Timex3 is timex tag for TempEval 2010. Various TIMEX3 [6] tags are: (a) One TIMEX3 tag used when (a1).there is no intervening token between temporal terms e.g. 8:00 pm. (a2) certain prepositions appear within temporal terms e.g. the

second of December. (b) Two TIMEX3 tag used when (b.1) Two temporal expressions that are ordered one relative to the other. They are joined $\rightarrow$ with prepositions and conjunctions e.g. 2days before yesterday. (b.2) two temporal expressions that can be related by temporal link e.g. 8:00 pm on Friday.

TEMEX tag has a various attributes, I used only (i) type: date, time, period or set. The format of the value attribute is determined by the type of TIMEX3.e.g a PERIOD must have a value that begins with the letter 'P' since durations represent a period of time. (ii) value: given as in ISO 8601 format.

Table 1. Markable Temporal Expressions

Type	Words
Always marked	
Pivot words- Nouns	□□□(day), □□□(year), □□□□□(week), □□□□□(month)
Pivot words- Day and Month Names	□□□□□□□(Friday), □□□□□(January), □□□□□(February)
Pivot Words- Known Date & Time formats	24 □□□□□'16, □□/□□□□□/□□□□, □:□□ □□□, □.□□ □□□
Festivals & Seasons	□□□□□□□□□□, □□□□□□(monsoon), □□□□□(winter),
Marked as part of temporal expression if they quantify a pivot word	
Quantifiers	
Hindi numerals and number words	□□(one), □□(ten), □(1), □(2)
Marked as part of temporal expression if they direct a pivot word	
Direction	□□□□(next), □□□□□(last), □□□□□(last),

words	□□□□(before)
Marked as part of post modifiers	
Other words	□□(at), □□(on), □□(of), □□, □□□□(about)

1) Temporal Expression Classification

Temporal Expression belongs to one of four classes: Date, Time, Period, Set. Date can be combination of day, month and year. e.g. (2015, 2 फरवरी), (छब्बीस जनवरी). Time can be in form of hh:mm:ss or hh:mm or (digit बजे) e.g. प्रात काल:11:00, सुबह 10 बजकर 18 मिनट. Period indicate any duration e.g. 9 साल और 8 महीने, ४८ घंटे, 18-24 जनवरी 2016. Set is time occur with some frequency e.g. हर महीने. Given examples are for Hindi language temporal information include in tweets. Normalization assign value and type to extracted temporal information as per timex3 tag.

Table 2. Various example Temporal Information

Date	Time	Set	Period
24 □□□□□'16	24.04□ □□	□□ □□□□	23 □□□□□ □□ 22 □□□□□
23 □□□□□ □□□ 1897	□□□□□□□□□ 04 □□□	□□□□□ □□□□	23 □□□□□, 1897 - 18 □□□□□, 1945
□□ □□□□□	12:12 PM IST	□□ 26 □□□□□	23 □□□□ 56 □□□□ □□ 4.091 □□□□□□
□□□□□ 26 □□□□□	18 □□□□□ □□ □□□□ 9 □□□	□□ □□□□□	12 □□ 14 □□ 2016
□□	13□□ □□□	□□□□ 5 □□ 8 □□□	□□□□□□□ 8 □□ □□
□□	□□□□ □□□□	□□ □□□	□□□□ 5 □□ 8 □□□
dd/mm/yyyy or yy mm/d d/yyyy or yy yyyy or yy/mm/dd yyyy or yy/dd/mm	□□□□ □□ □□□□	□□ □□□	□□□□□□ □□ □□ □□□□□ 3 : 31

2) Temporal Expression Normalization

Normalization is a process to give norm value of given temporal expression. Normalizer provide the TimeML type and value attributes for each temporal expression given in input. This program takes a temporal expression and returns the normalised type, value and modifier according to i2b2 2012 scheme (derived from TimeML).

Table 3. Type and Value attribute example

Temporal expression	Value	Type
Eleven in the morning	T11	Time
□□□□□□□□ □□□	2016-03-04TEV	Time
eigth of march	2016-03-08	Date
6 □□□□□	P4M	Period
56 □□□□□	PT20M	Period
□□□□□□□□□□ 04 □□□	T04:00	Time
□□□□□ □□□□□□	2016-03-07	Date
26 □□□□□□	2016-01-26	Date
□□□□□□□□□	P1	SET
□□ □□□	P1	SET

4 IETE Fire 2016 Corpus

Existing ILTIMEX2012 corpus [11] is made of news articles from the Hindi newspaper Navbharat Times. This corpus data are written in well-formed grammar specification.

The IETE FIRE 2016 corpus is constructed using Hindi tweets for general domain. Manual tagging of the chosen files was done. Table 5 contain dataset statistics IETE FIRE 2016 corpus. Table 6 highlights a comparison between the ILTIMEX2012 Corpus and IETE FIRE 2016 corpus.

Table 4. Dataset: statistics

Dataset	No of tweets	No of tokens
Train	3500	32k
Test	500	4.5k

Table 5. comparison of ILTIMEX2012 and IETE FIRE 2016 corpus

Detail	ILTIMEX2012	IETE FIRE 2016
Nature of documents	news	Tweets
Dataset statistics	500 documents	4000tweets
Total number of tokens	About 150k	About 36.5k
Total number of temporal expression	1840	2500

5 Proposed solutions

5.1 Rule base approach

Rule base approach [12] is based on set of hand crafted rules for temporal information identification. These rules are basically regular expressions and apply to plain input files. In rule-based expert systems knowledge is represented in an IF-THEN form: IF <set of conditions> THEN <decision>. A limited subset of natural

language - supplemented by specified relations and operators - is used to formulate the rules. Rules are well structured documents for an information retrieval system, particularly since the number of rules in a rule-based system remains manageable.

Table 6. Identification Rules: Statistic

Temporal class	No of Rules
Date	33
Time	40
Period	16
Set	8

Table 7. Normalization Rules: Statistic

Temporal class	No of Rules
Date, Time, Period and Set	37

5.2 Conditional Random Field approach

A set of hand crafted rules are required in rule based approach. If number of iterations for manual addition of rules increase, rule base approach can be a costly.

CRF [5] takes the neighboring samples or the context information into consideration. In case of temporal information recognition problem, document of text or a token of a sentence is taken as an observation sequence and the label sequence is taken as state sequence. In CRF approach system tokenize whole dataset and extract 36 features. These features belongs to following categories.

Morpho-lexical: This category includes the token, its lemma. For lemma and POS tags I use Hindi POS tagger by TDIL India. Boolean features contain if word is lower case, upper case, number. To handle temporal expressions phonetic features: for seasons, past references (e.g. पिछले, पीछली), future references (e.g. अगले), prepositions, conjunctions and adjectives. Total 30 morpho-lexical features have been engineered.

Gazetteers: I used gazetteers: Month name (e.g. सितंबर, सितं, August, June, जूनवरी), Day name (e.g. सोमवार, Monday, Sunday), Festival name (e.g. गणतंत्रदिवस) and

other which contain word (e.g. रात, पिछले, वर्ष, आज, दोपहर, today, morning). Total 4 gazetteer-base features have been engineered.

Syntactic: Chunks and propositional noun phrases are syntactic features. This achieved using Shallow Parser by TDIL India. **See Table: 8**

5.3 Support Vector Machine approach

The main work of this approach [5] is to find whether a specific vector belongs to a particular target class or not. In this approach, the training file data and the testing file data goes to the single dimension vector space. During training, a hyper plane is generated that is used to classify the members into two classes: positive and negative classes that happens on the opposite sides of a hyper plane. This approach calculates margin.

FEATURE is used to change the feature sets (window-size) for chunking. Feature generated in Yamcha should be written in the following format:

Features: -2:0 -2:1 -2:2 -1:0 -1:1 -1:2 0:0
0:1 0:2 1:0 1:1 1:2 2:0 2:1 2:2

Tag_Features: -2 -1

6 Result & Discussion

The systems were made to tag the plain testing files and the output files were compared with the corresponding

manually annotated files. Table 21 highlights the experimental results for the above discussed approaches.

Table 8. Context features

POSInfo of token at i-2, i-1, i+1, i+2
Chunkinfo of token at i-2, i-1, i+1, i+2
isPivot of token at i-2, i-1, i+1, i+2
containsDigitsAndPunc of token at i-2, i-1, i+1, i+2
bigram feature - POSInfo of token at i-i and POSInfo of current token
bigram feature - POSInfo of current token and POSInfo of token at i+1

Table 9. Evaluation of three approach

	SVM		CRF		Rule Base		Normalization
	ID	ID+CL	ID	ID+CL	ID	ID+CL	
Manual Tagged	720	720	720	720	720	720	720
Total Tagged	675	675	570	570	665	665	728
Correct	494	498	533	529	585	585	574
Missing	51	47	151	151	63	63	47
False positive	175	175	37	41	125	130	99
Strict Evaluation							
Precision	0.90	0.91	0.93	0.92	0.82	0.81	0.85
Recall	0.73	0.73	0.77	0.77	0.90	0.90	0.92
F1-measure	0.82	0.81	0.85	0.84	0.86	0.85	0.88

ID= identification, ID+CL= identification and classification,

6.1 Analysis

Analysis of evaluation of the developed systems is critical and is presented as follows. The focus of the analysis is on the missing entries observed during the evaluation. When a temporal information is wholly missed by the system, missing temporal patterns get are generated.

List of missing temporal patterns in Identification phase:

In SVM approach

When two type of temporal patterns are conjuncted, they identified as separate patterns instead of one pattern. In

given tweet, 8 identified as 8 and two separate patterns.

- Both in SVM & CRF approach
- When temporal patterns includes festival names
- When Two month name are joined with conjunctions are missed.
- Tweet: #OddEvenFormula -
 they identified without prefix. In given tweet 20 identified as 20, not as whole entity 20.
- Tweet: 20 ...

In Rule Base approach

- When pattern occur like 1990, 2015 without any prefix or suffix word, are missed.
- When pattern include two number with word e.g. 21 23, this create ambiguity.
- Tweet1: 21 23
- Patterns are not identified because rules to identify these patterns are not present in rule base due to absence of these patterns in training data. e.g., 05:20 To 06:05.
- Tweet1: @_memoriespain: ,
- Some Temporal patterns needs so complex rules, are missed. e.g. 23 56 4.091; 12 ' ' 23

□□□□ '□□□□ □□□□ □□ □□□□'; □□ □□, □□ □□□□, □□ □□; 23 □□□□, 2016 □□ □□□□□□ □□□□ □□□□□□□□ 01 □□□□□□, 2016

In three approach

- When temporal patterns contain conjuncted parts e.g. 19 □□ 20 □□□□ then 20 □□□□ only identified. 22 □□ missed in identification. Other example: 18,19,20 □□□□
- Temporal patterns e.g 5-5, 19/4 due to ambiguity. In tweet1 11-2 indicate run and in tweet2 16/11 as part of address, in tweet3 8/4 as date.
- Tweet1: □□□□-□□□□ □□□□ 5-2□□.
- Time expressions e.g 10:03, □□□ 8.3, □□□□□□ □□ □□□ are missed because of generally patterns occurs with suffix. e.g. 10:03 □□□, □□□ 8.3 □□□
- Tweet: □□□□□□ □□ □□□ □□□□□□□□ □□□ □□□□□□□□ 19.01. □□ □□□□ □□□□ □□□ 8.5 MGD UGR □□□□ □□ □□□□□□□□□□
- Some temporal patterns like festival names written in transliterated form that are missed.
- Tweet: □□□□□ □□□□ – □□□□□□□□ □□ □□ □□□□□□ □□ □□□ □□□□□□□□ □□ □□□□□□ □□□□□□ □□ □□ □□ #Airlift □□ 4 □□□ □□□ 50 □□□□□ □□□ □□□□
- Due to absence of some temporal expressions format in training data, missed in identification. e.g. □□ □□□, □□ 10□□□ □□□, □□□□, □□□ □□, □□□□-□□□
- Twitter user post tweets without any grammatical and any format. Therefore possible that temporal pattern format occur currently may not be as per training data. e.g. 26 □□□□□□□□' 2016, □□-□□□□□ □□□□.

List of missing temporal patterns in Normalization phase:

- When temporal entity occur without space are missed. e.g. 15□□□□1947
- When two temporal expressions joined with conjunctions, are missed. e.g. □□□□□□ 8 □□ □□; □□□□ 8-14 □□ □□□□□□ 8-09; □□ - 20 □□□□□□ 2016; □□□□-□□□□.
- When temporal expression contains words like □□□□□□□□, □□□□□□□□ missed. e.g. □□□□□□□□ □□ □□□ 5:30 ;□□□□□□□□□□ □□ □□□□□□ □.□□;
- Some temporal words e.g. □□ create ambiguity in meaning. □□ may be as tomorrow or yesterday.
- Date pattern occur with dd mm yy format missed. e.g. □□ □□□□□□ □□
- Some temporal patterns missed. e.g. □□ □□□□□□, □□ 10□□□□ □□□□, □□□□
- Some Temporal patterns needs so complex rules, are missed. e.g. □□ □□□, □□ □□□□□□, □□□□ □□□; 23 □□□□□□, 2016 □□ □□□□□□□□ □□□□ □□□□□□□□ 01 □□□□□□□□, 2016.

CONCLUSION

System uses different approaches to temporal expression identification and classification for Hindi Tweets. Probabilistic model (CRF) is better than Vector based model (SVM). For Identification phase, Rule base give best result among all three proposed approach: A rule base approach has F1-measure: 0.85 and two statistical approaches has F1-measure: 0.83(CRF) and 0.71 (SVM). For Normalization phase Rule base approach give F1-measure 0.83.

REFERENCES

1. Angel X. Chang, Christopher D. Manning. "SUTIME: A Library for Recognizing and normalizing Time Expressions". In eighth international conference on Language resources and Evaluation (LREC) 2012.

2. Hector Llorens, Leon Derczynski, Robert Gaizauskas, Estela Saquete. "TIMEN: An Open Temporal Expression Normalisation Resource". In eighth international conference on Language Resources and Evaluation (LREC) 2012.
3. J Strötgen, M Gertz. "A Baseline Temporal Tagger for all Languages". Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 541–547.
4. J Strötgen, M Gertz. "HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions" Published in Proceeding SemEval '10 Proceedings of the 5th International Workshop on Semantic Evaluation Pages 321-324.
5. John D. Laerty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, pages 282289, 2001
6. Michele Filannino, Gavin Brown, GoranNenadic. "ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge". In June 2012.
7. N Ramrakhiyani, P Majumder. "Approaches to Temporal Expression Recognition in Hindi". In Journal ACM Transactions on Asian and Low-Resource Language Information Processing Volume 14 Issue 1, January 2015.
8. NattiyaKanhabua ,Sara Romano, Avaré Stewart. "Identifying Relevant Temporal Expressions for Real-World Events". In 35th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR) 2012.
9. NattiyaKanhabua, Sara Romano, Avaré Stewart, Wolfgang Nejdl. "Supporting Temporal Analytics for Health-Related Events in Microblogs". Proceeding CIKM '12 Proceedings of the 21st ACM international conference on Information and knowledge management Pages 2686-2688.

10. NaushadUzZaman, Hector Llorens, James Allen. "TempEval-3: Evaluating Events,

Time Expressions, and Temporal Relations" represented in March 2012.

11. NitinRamrakhiyani, P. Majumder. "Temporal Expression Recognition in Hindi", Volume 8284 of the series Lecture Notes in Computer Science pp 740-750.
12. NitinRamrakhiyani, P. Majumder. "A Regular Expression Based Time Expression Recognizer in Hindi", presented in 2013.
13. Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., and Radev, D. 2003. TimeML: Robust specification of event and temporal expressions in text. In New Directions in Question Answering, 28–34