International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Event -Time Relation in Natural Language Text

Vanitha Guda, Suresh Kumar Sanampudi

Abstract*: Due to the numerous information needs, retrieval of events from a given natural language text is inevitable. In natural language processing(NLP), "Events" are situations, occurrences, real-world entities or facts. Extraction of events and arranging them on a timeline is helpful in various NLP applications like building the summary of news articles, processing health records, and Question Answering System (QA) systems. This paper presents a framework for identifying the events and times from a given document and representing them using a graph data structure. As a result, a graph is derived to show event-time relationships in the given text. Events form the nodes in a graph, and edges represent the temporal relations among the nodes. Time of an event occurrence exists in two forms namely qualitative (like before, after, during, etc.) and quantitative (exact time points/periods). To build the event-time-event structure quantitative time is normalized to qualitative form. Thus obtained temporal information is used to label the edges among the events. Data set released in the shared task EvTExtract of (Forum for Information Retrieval Extraction) FIRE 2018 conference is identified to evaluate the framework. Precision and recall are used as evaluation metrics to access the performance of the work with other methods mentioned in state of the art with 85% of accuracy and 90% of precision.

***Index Terms*: Natural Language Processing, Events, Times, Event-Time Graph, Temporal Question Answering.**

I. INTRODUCTION

In the present digital information era, there is exploitation of data which is results to huge number of textual type of sources (e.g., web data, Tweets, data generated from news, history records, medical type records, legal documents) these all categories of generated texts are mostly with the descriptions of events, extracting and analyzing events from a given document is found to be an essential task. In linguistic terms events in the text are referred to as event mentions. In real scenario most part of the text will have the ambiguities, incompleteness with that the representation of real-world events and representing the relations is crucial task.Mostly sentence level event extraction works carried out by the existing works of NLP, document level events detection done within the mentioned topic or specific topic. In the Topic Detection and Tracking (TDT), "event" is defined to be a happening or something that occurs at a specific time and place, topic can be represented as the in the form of news stories associated with the facts as stated in [1]. In Topic detection it mostly depends on vector space model as stated in [2]. The most recent methods of vector space models are based on entity classes (e.g., named entities, noun phrases, collocations, these entities are important because often identifies the participants of an event as stated in [3] [4] [5].

Event-oriented information needs to involve structured text or queries rather than depending on keyword type (e.g.'What are the countries that Prime Minister Narendra Modi has visited and in which of his visit he declared protection laws?'). This type of query representation most useful in building of temporal question answering (QA) systems as stated in [6] e.g.'who won both the Hyderabad and Vizag marathons three years in a row?'). The QA system results in multiple responses, where the gap of missing and incorrect information is filled by merging the data obtained from multiple sources.

In this paper we are presenting a framework for events, times extraction in given NLP and represents the event-time relations with the help of a graph data structure. Graph with nodes edges, in event graph structure nodes are the events and relation between nodes represents the temporal relations that are the edges and finally its looks like a labelled graph. The initial stage of framework extracts event-time relations at sentence-level events, later event-time relations are inferred for document-level events.

The main contribution presented in the paper talks about the development of a framework for constructing event-time graphs. This event-time modeling is language-independent. Paper focuses on the processing of English language texts, state-of-the-art methods are compared to evaluate the performance. Forming event-time structures or deriving graphs from the given text is observed to be a challenging task due to several reasons majorly following are the reasons

Machine learning and rule-based methods used tasks like dataset understanding, model building and finally evaluation.

- ✓ To do the pipeline the activities like rsing, named entity recognition, entity co-reference resolution.

The rest of the paper explains as follows. Section 2 gives the key points of the related work in event extraction and also time extraction from a given text. Section 3, explains the proposed framework.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Vanitha Guda*, Department of CSE, Chaithanya Bharathi Institute of Technology (A) Gandipet, Hyderabad , Telangana, India.
Dr S.Suresh Kumar, IT Department, JNTUCEJ, Nachupally Jagityal Karimnagar,India.

Next 4th-section elaborates the process of building the event-time graph. Section 5 illustrates the corpora used for experimentation. Section-6 presents the evaluation of outputs obtained for the proposed framework. Finally, the outline conclusions and future scope in section 7.

II. RELATED WORK

Events -Time are important terms in most of the NLP applications like text summarization, question answering systems. The available literature talks about the extraction of events and times as independent units in the system. In 2.1 section gives a brief overview in event extraction and time extraction explained in next Section 2.2.

2.1 Event Extraction

The events extraction concept is also recognized as the step in Message Understanding Conferences (MUC) [7], the structure of events are related to particular fields or domains. The structures were predetermined, and the task was to identify the events and classifies them. The initial methods/ procedures and the first systems of events extraction are rule-based [8]. For each domain, new extraction patterns are designed as stated in [9]. The newly systems and methods are focusing on obtains templates automatically, from the study conducted on even extraction are those works can be differentiates in three classes, (1) Traditional Data driven methods (2) Pattern type or knowledge-driven method and (3) Combination of data and knowledge as hybrid methods.

i. Data-Driven approaches: Data-driven approach generates the rule for event extraction**.** Quantitative methods and lexical features are used to derive template based rules. It encompasses all quantitative approaches to automate language processing. As stated in TempEval-2013[10] (Temporal Evaluation), Zhen[11], STEP [12], explain the methods and features used to generate the rules for extracting the events. Thought these methods are simple to implement; they do not consider semantics during rule construction. Rules are not exhaustive, need to build a new rule when a new situation arises. All the methods used in this class are non-knowledge based.

ii. Knowledge-driven: This type of method uses knowledge about the content and encodes them into patterns. For the generation of a pattern, rules are constructed using lexico-syntactic or lexico-semantic features. Lexico-syntactic exploits use syntactical information. Lexico-Semantic uses the meaning of the information along with lexical representations. As stated in REES [9], EVITA [13] explain the methods which make use of syntactic and semantic information in building the event-patterns. Though pattern-based approaches have the advantage of using very less amount of data, patterns definition require prior domain knowledge. Designing and maintaining patterns is difficult when patterns need to be scaled-up to cover more situations. The feasibility and accuracy of rules depending on the user's knowledge of linguistics and domain expertise.

iii. Hybrid method: This method takes the advantages of data-driven and knowledge-driven methods. Besides, hybrid methods use machine learning approaches. In this method, expert knowledge is applied to the output of a statistical model to prune unwanted results or to include information that could have been missed by statistical models. Also, researchers have combined statistical approaches with (lexical) knowledge, where encoded lexical knowledge forms the features for supervised learning. As stated in FrameNet [14] or VerbNet [14] , methods are build using hybrid methods. In these methods, complexity increases due to increasing of data and the combination of the multiple techniques. Expert knowledge is needed for efficient event discovery because most of these systems are domain specific.

2.2 Time Extraction

Time is a temporal expression that indicates the point or period of happening in a situation. The representation of time exists in two forms, Quantitative form (15th July 1985, 15/07/1985) and Qualitative form (After the break, during the war). The works for time extraction can be categorization into rule-based and annotation systems.

i. Rule-Based Systems: Rule-Based systems build templates on labeled corpora such as [15] these templates recognize quantitative times which are calendric times. Later GUTime: a Perl temporal tagger provided by Georgetown University which is based on TempEx consisting of components for the extraction of events, temporal expressions, and temporal relations for recognizing and normalizing temporal expressions in English text implemented for the clinical narratives. To extract and normalize time expressions detection and normalization Temporal tagger it was adapted from publicly available Temporal Tagger. Next is BIO classifier [31] tags each input token as either Beginning, In, or Out of a time expression with nine tags. Temporal Expression Recognition as a BIO token-chunking task, where each token in the text is classified as being at the B(eginning) of, I(nside) of, or O(utside) of time expression. A rule-based system is only for the Quantitative times and Domain Specific.

ii. Annotation Based Systems: Annotation for time expression analyzes the ISO times and builds mark-up language is a Markup or representation language used to capture temporal information but it can only represent the language cannot perform reasoning of the generated events with times. The standard annotations for event and time information settled by TIMEML [21], and for the dataset are available in TimeBank [20] quoted in the literature. In ACE[22] event extraction explained with templates in robust manner with arguments and event co-reference resolution. TempEval campaign[23] explains the structures between events-time and also event-time with DOT means document creation time.In the next TempEval second campaign [21], introduces about the event anchors and time expression extraction, and in the given sentence finding the relations between events and time. As a task of NLP i2b2 challenge identified to work by using the records of clinical work [24] most of the same job of TempEval-2. In observations of the present systems, events and times were designed to use for a particular domain (financial, medical),the Semantic nature and representation of time is missing in temporal extraction methods.

III. WORK FLOW OF THE PROPOSED METHOD

This section presents a workflow of proposed framework building event-time graphs.

The text or the documents are considered to be the input of the work, the output of the system is considered to be the event time graphs. In the first step text document is fed as input then processed through basic NLP pre-processing steps such as Tokenization, lemmatization, stemming, parts of speech tagging, parsing and Named entity recognition. Thus processed document will be given for extraction of events and times. After the extraction of events and times, the relation between events and times need to be identified for the construction of the event-time graph. The framework proposed is shown in figure-1.

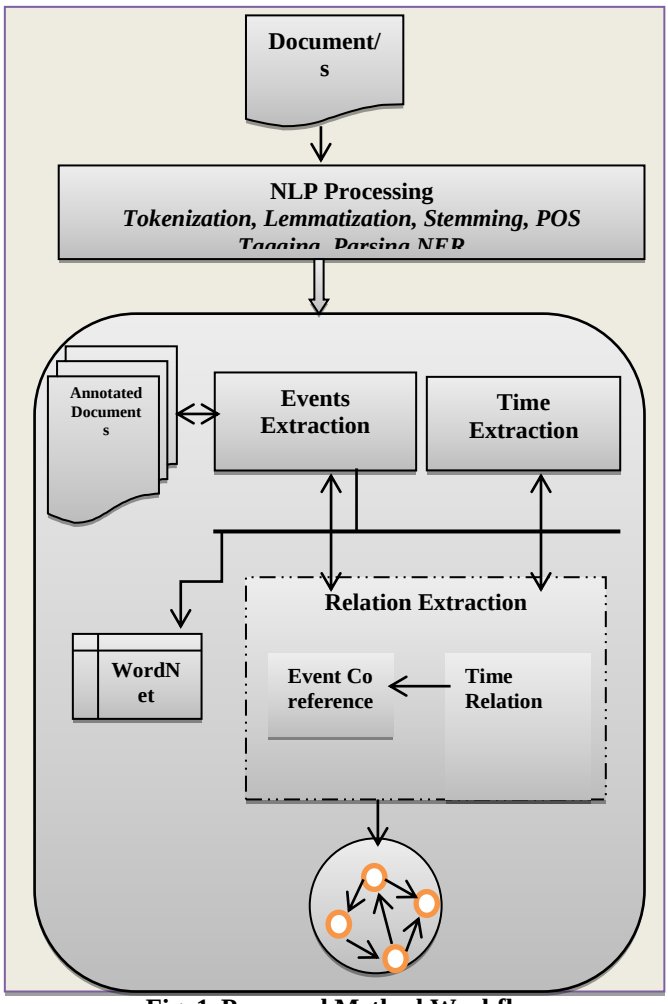

Fig. 1. Proposed Method Workflow

3.1 NLP Pre-Processing: Basic NLP pre-processing involves the following steps:

- **Abbreviation Expansion:** if the text involves any contractions need to expand, generally these are shortened version of words or syllables. Most of the contractions are exist in English text in written or spoken forms. English contractions exist in the text that word can be expanded by adding or removing one of the vowels of that word. For example, don't to "do not" and I'd to "I would". Converting each contraction to its expanded, original form helps for text standardization.
- **Special Characters Removal:** these are symbols or specific characters basically non-alphanumeric characters or even numeric characters (depending on the problem), by using regular expressions (regexes) special symbols can be eliminated from the text.
- **Tokenization:** the process tokenizing means paragraph splitting into sentences and sentences into words or individual words or punctuation through a similar process. Most commonly this split across through the white spaces.
- **Stop words elimination:** Stop words can be articles, conjunctions, prepositions namely a, an, the, etc. These do not form any significance so they are removed from the text.
- **Lemmatization:** It removes word's affixes to get the base form of a word. Main root word can be in base form but not as the root stem. Means the root word is always explains the lexical word, but the root stem is not lexical. Root word can also state as lemma and that will be in the dictionary of the language.
- **Parts of speech Tagging (POS):** Syntax of the token can be obtained by POS tagging to tag the word in the text with its syntax or POS. These syntax tags help to detect the syntax features of the tokens.

For example, consider the following text as input to the proposed framework.

Original Text: *Nah I don't think he goes to USA, he lives around here though.*

After performing each of the NLP-preprocessing steps the intermediate outputs at each stage are shown below. NLP preprocessing can be with the help as stated in [29].

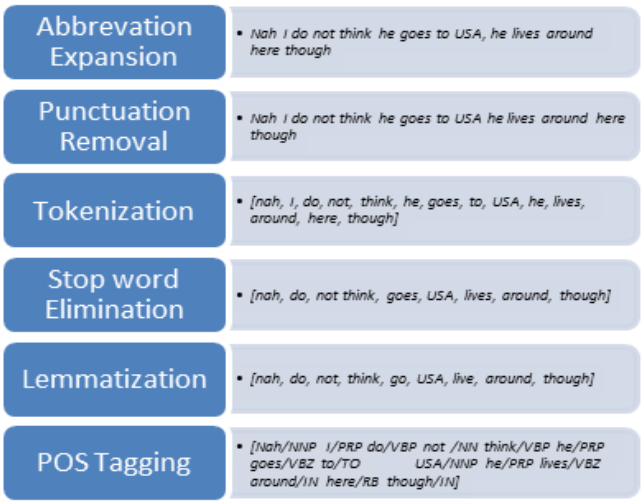

Fig. 2. Steps of stages in Preprocessing

3.2 Event Extraction

After the basic NLP pre-processing, the POS tagged tokens from the text will be the input for the event and time extraction component. In English text, events are derived using syntactic and semantic features. An approach is built where the nouns and verbs are extracted. At the initial stage, all the verbs are treated as events. In the next stage, the words that have the noun and verb tags are resolved from ambiguity to identify the nonverbal events.

Given a POS-tagged text as input a triplet is constructed **(T, V_t, N_t)** where T is the token itself which occurs in the text, V_t is the verb form of the token and N_t is the noun form of the token. Events are referred to be the action forming verbs.

In some cases, even the noun forms the events but in this rule based systems it is difficult to resolve the ambiguous cases of non-verbal events, like for example word "declare" can be identified as a verb by the POS tag but it can be a noun also. To identify these cases machine learning approaches are used to predict the semantic role of an event. Machine learning approaches like Conditional Random Field (CRF), and Semantic Role Label (SRL), WordNet are used to resolve nonverbal events.

Conditional Random Field (CRF) is one of the discriminative classifiers which correspond to conditionally trained the probabilistic model. The conditional probability of a state sequence X = {x1,x2,..xT} given an observation sequence Y= {y1,y2,.,yT} function $f_k(x_{t-1}, x_t, y, t)$ is a feature function whose weight k is to be learned via training. CRF considers the tense and aspect of a lexicon for event detection. The main advantage of using CRF is that it has sequential data handling after recognition of one token the previous context of that token can be taken into consideration to calculate the conditional probability. But still, some nonverbal lexicons were not resolved as events.

To overcome the problem Semantic Role Labelling (SRL) WordNet are used to detect nonverbal events. SRL retrieves the constituents and its arguments, adjuncts of a lexicon also. It the one of the process to get the constitutes along with the suffixes (-ed, or, ee, er). Generally the nominalizations often the verbs, if nominalizations can also are derived as nouns, these are morphological way considered from verbs and nouns because of its suffixation. For example, consider a sentence

Eg: "*All parties were declared inspected to the satisfaction of the investigation team and with the full cooperation of Mumbai authorities, Ram said".*

SRL output for the sentence is as follows:

[ARG1-All sites] were [TARGET declared] to the satisfaction of the investigation team and with the full cooperation of Mumbai authorities, [said].

In the first iteration of the sentence, the word *"Declared" is an* event. In Next iteration, "said" is identified as an event. The retrieved target words are considered as event words, it is marked as many of the target words are represented as the event expressions but that are exist many nominalized event expressions (i.e., deverbal nouns) that are not identified as events by the supervised CRF. These nominalized expressions are correctly identified as events by SRL. Even after identifying semantic roles still, there are some cases where noun-verb ambiguity exists, for some of the words like 'war', 'attempt', 'tour'. To overcome this WordNet is attached to SRL. Wordnet checks each words if any one of the word senses as a verb then it will be announced as event and if the token appears as noun and verb then also it is treated as an event. (Eg: "Declared" is a word the stem word is "declare"). The WordNet features can be used to extract different lexical categories, such as part-of-speech (POS), stem, hypernym, meronym, distance, and common parents, and demonstrated its worth in many tasks. The performance improvement explained in results section.

For example, given a text

Dr. Abdul Kalam is known as the 'missile man', born on 15th October 1931 in Tamil Nadu. Kalam graduated in science from St. Joseph's College, 1954. He reports every day in lab his aim is vision space.

Step1: The output of the NLP pre-processing

Dr/NNP Abdul/NNP Kalam/NNP known/NN as/IN the/DT â/NNP €˜/NNP missile/NN manâ/-NONE- €™,/. born/NN on/IN15th/CDOctober/NNP1931/CDin/IN Rameswaram/NNP,/, Tamil/NNP Nadu/NNP. ,/, Kalam/NNP graduated/VBD in/IN science/NN from/IN St/NNP./. Joseph/NNP '/POS s/NNS College/NNP/1954/CD. Kalam / NNP reports/NN in/IN lab/NN aim/NN vision/NN space/NNP.

Step2: The output of the POS tagger

Dr/NNP Abdul/NNP Kalam/NNP known/NN missile/NN manâ/-NONE- €™,/. born/NN Rameswaram/NNP Tamil/NNP Nadu/NNP graduated/VBD science/NN St/NNP Joseph/NNP College/NNP/ reports/NN lab/NN aim/NN vision/NN space/NNP.

Step3: After passing trough conditional random field, SRL with wordnet the following words are the output i.e, extracted as events.

"Missile"," born", " graduated ", "science", " report"," aim".

The output of the Event Extraction task of the framework extracts the events by using the syntactic and semantic feature from a given natural language text irrespective the domain. Integration of CRF, SRL, wordnet with hand-coded rules addressed the gaps identified in the existing events extraction systems.

3.3 Time Extraction:

Time in the NLP text exists in quantitative and qualitative forms. Several methods were developed in the past to extract various forms of time expressions. In this framework quantitative time expression i.e. calendric times mentioned with specific date time in standard ISO format were directly recognized by using SuTime. Time expressions like "independence day", "Mother's day" are not captured by SuTime. To extract such temporal expressions, pattern-based rules are developed and integrated into the existing framework. Rules are developed for calendric holidays of INDIAN scenario. Below are some set of rules from a holiday package.

//Samplerules

{/new//year/$POSS?/day/?)=>IsoDate(NIL,1,1)//January1st
{/republic/ /day/)=> IsoDate(NIL, 1, 26)}// RepublicDay
Day {(/independence//day/)=>IsoDate(NIL, 8, 15) //Independence Day

Extraction of Time Expressions: Quantitative time expressions are in a numeric form where qualitative forms are not. Time expressions are obtained in lexical, syntax and semantic features.

Quantitative time expression as lexical features these are calendric times mentioned with specific date time in standard ISO format these quantitative time directly recognized by using SuTime.

- **Qualitative time** expression as a syntactic feature of the time represented with Allen's algebra with 13forms of relations (Before, after, during, overlap) between events these are not directly specified with a number. In time expression representation semantic feature of time considers specific holiday events like (independence day, mother's day, etc.,) these semantic representations of time not captured by SUTime, for that we implemented pattern-based rules as added as holiday package to SuTime.

- SuTime doesn't capture Qualitative time 13 relations of Allen's and SuTime does not identify semantic time relations like Indian holidays. In our work by using lexical, syntax features es differentiating a number as a numeric value and a time expression with time relation. By adding holiday package to the SuTime where it will extract time in semantic form like *(Independence Day, Mother's day of Indian calendric holidays).*

❖ The overall process of time expression first lexical lookup feature means mapping names to numbers, units to ISO values, etc. Next is context-dependent classification: determining whether the time is a point or duration, looks forward or backward, makes specific or generic reference with the help of lexical, syntax and semantic features. The reference time for time expressions whose values must be computed in final computation combining the results of all of these steps to produce a final normalized value to detect time expression

For Example the given Text:

Kalam was born on 15th October 1954. On Independence day discussed kalam's lifehistory. He visited Delhi after 26th January.

The output after the tokenization:

Kalam//NNP was//VBD born//VBN on//PP 15th//JJ October//NNP , 1954/CD. On //IN Independence//NNP day//NN discussed//VBD kalam//NN life //NN history//NN. He//PRP visited//VBD delhi//NN after //Prp 26th//JJ January//NNP

Output after the time Expression recognition:

15th October 1954 Independence Day After 26th January/ //"15th October" "1954" "26th January" are Quantitative time, Independence day semantic time After is qualitative time.

3.4. Relation Extraction: In this step, the relationship needs to establish between the extracted events and time expressions. Event reference relation and time relation declares the final events and time expressions with the lexical, syntax and semantic features and that are useful for event-time graph construction from the extraction modules.

The rules to declare the events and time from the given text by the following rules:

- **Suffix rule**: In text Deverbal nouns are usually identified by the suffixes like '-tion', '- ion', ' -ing' and '- ed' etc.
- **Noun and Verb Rule:** Noun and verb combinations are searched in the sentences of the test set. The non-NE noun word tokens are considered as events.
- **Nominal and Non Verbal Rule**: Nominals and non-de verbals can be identified by the complements of aspectual PPs headed by prepositions like during, after and before, and complex prepositions such as at the end of and at the beginning of, etc. The next token(s) appearing after these clue word(s)/phrase(s) are considered as events.
- **Event Noun Semrule**: Some expressions like the occurrence of any period of are most probably the event nouns.
- **Object Rule**: the word like objects of aspectual and time-related verbs, such as *"have begun a campaign" or "have carried out a campaign".* The notion of Named Entity is that appear after the
- Mapping of the rules to declare the time events by the following rules: The events and time expressions are obtained from the above individual steps next is to the map the events which associate with time expressions that are said to be temporal expressions or time expression. In these events which relates to other events can be noted with qualitative time expressions, that will be the relations between events.

IV. EVENT-GRAPH

In Present work we are defining that the event- graph structure in labeled edges with events and times in a mixed graph manner, the vertices represent events, edges represent the relation between events with time expression. The graph represented as G in a notation of tuple that is V vertices, E edges , S directed edges , m is mapping and r is the relation that can be finally stated as $G = (V, E, S, m, r)$, Event from 'e' can be assigned to a type (e.g., Occurrence, happening or Reporting) and consists of features of event and its arguments.

i. In this work, events are considered as relations and that denoted by 'e the 'r' is relations based on the following:
ii. Fact Feature of the events: in real-world events represents facts, Eg: in question answering for factual event is (e.g., 'Who killed MahatmaGandhi?'), and nonfactual event (e.g., 'Who did not win a prize in hackathon 2016 ?' or 'When might Clinton resign?'). In our work we didn't focus nonfactual events or hypothetical (e.g., 'he can win'), future (e.g., 'He will win'), negative (e.g., 'They did not win'), and event mentions ('If he had won '), Since we aim to represent factual event are real-world events which actually occurred, and that may important for the ordering of the events (Karttunen and Zaenen 2005) because the events can be pointed on a timeline.
iii. Event features consisting of Token features: basic word, the stem, lemma and POS tag.
iv. iii.Context features: for tokens grammatical syntax verification and syntax relations of a token, the type of chunk (e.g., NP, VP, PP) and chunk's adjacent are identified in this.Features-Modifier: these are to differentiate the factual, non-factual category of events.
v. General features: indicates most of the words are capitalized or not (named events,

such as 'World Cup' are capitalized).
vi. Event argument types: In extraction to get the robustness, we considered event arguments in four categories, that like an agent, location, time and target. These types of arguments are suitable to answer four main wh-questions: 'where, and when, Who did what to whom,
vii. Relations Type: Semantic may exist between events, temporal, causal with semantic relations also. In a document event references graph presented in Figure-3.

Input text for the graph: "The boat was pushed on the rocks. A wave swept four crew members Stallion rounded Farallon Island Stallion allowed Chase a head start Stallion raced on Sunday, Stallion raced around Farallon Island around Stallion competed against Chase. The boat sunk on the shore 'The boat which won the race. The crewmen storm swept'.

V. ABOUT THE DATA SETS

In our work we used the corpus of Forum for Information Retrieval (FIRE-2018) Datasets contains various semi-structured documents (XML) consisting of events related to 11 categories of files. The categories emphasized are Accidents, Crime, Cyclone, Earthquake, Fire, Floods, Shootout, Storm, Suicide Attack, and Volcano.

By nature, the document is in English language and not domain specific, After training the Model with FIRE[27] dataset we faced an issue, with fewer data. This less data isn't helping to train the model accurately. Thus we started scraping news articles from news wires and social media. We created a News data set which falls under the same 11 categories.

// Sample document from the newly formed dataset:
***"A** 23-year-old man was taken to Rockhampton Hospital after being trapped inside his car. His injuries were not life-threatening. The 52-year-old driver of the vehicle and two other people were not injured. Traffic diversions have been put in place until the scene is cleared. A seven-year-old boy was taken to Theodore Hospital for spinal precautions after a two-vehicle crash near Dingo earlier this morning."*

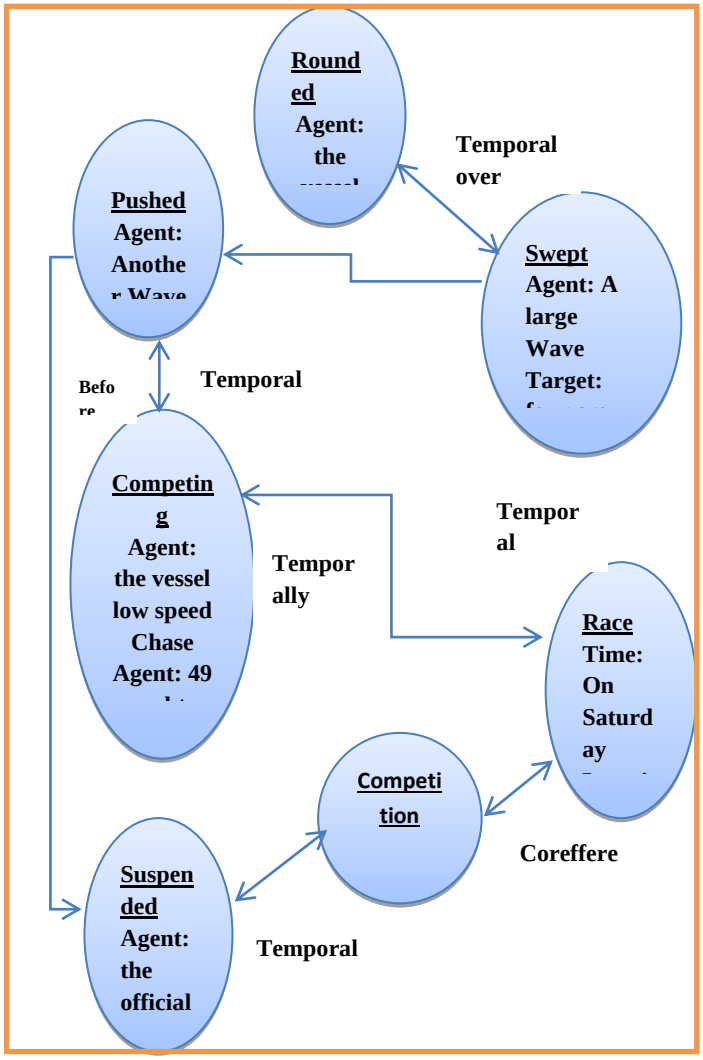

Fig.3. Example for event time relations graph from a news story from news corpus

The total we created 200 documents of news because news data contains more event information. In this, each document annotations are removed and 160 used for training and 40 documents for the test set. In that 160 documents also through cross-validation considered for train and test in 10 cross folds, and measures are precision-recall f-measure the accuracy applied for the various classifiers, the results section presents the values of each step.We used a python script to change the documents from XML format and strip them into a format suitable for our model. Basically, the script checks for any XML tags in each document and removes the tags from our document. Thus the final text in the document will be plain text without any XML tags.

VI. EVALUATION OUTPUTS AND DISCUSSIONS

6.1 Evaluation of Event Extraction

For the evaluation purpose SemEval [28], TempEVal and MUC data sets are considered. Both SemEval and MUC's data and data sets are text documents all are not related to any field or domain. Metric used to measure the accuracy are F-measure which computed with relevance measures of precision and recall.

The evaluated results of MUC [30] contain two hundred sentences over the documents and consist of verbal and non-verbal events presented in Table-1. Performance of F-Measure increases 3% from base hand-coded rules compared with last method of hybrid one. Table-2 presents noticeable improvement (i.e., 4%) by using CRF along with WordNet. With composite method the rules may increases F-measure of 5% improvement in comparison with previous used methods. Calculated accuracy methods with each method obtained results are explained in the tables.

Table 1. Results for Event Extractin MUC DataExample table

Handcoded Rules	**63.32**	**65.15**	**65.18**
CRF	**67.21**	**68.12**	**68.64**
CRF+ WordNet	**70.2**	**72.21**	**69.18**
Hybrid (or) CompositeRule	**72.21**	**73.38**	**73.74**

Table.2. Results for Event Extractin MUC DataExample table

Method	*Precision*	*Recall*	*F-measure*
Handcoded Rules	**75.32**	**77.14**	**76.21**
CRF	77.21	78.91	78.05
CRF+ WordNet	**79.2**	**81.22**	**80.19**
Our Event Extraction Model	81.11	84.23	82.64

6.2 Evaluation of Times Extraction

To obtain the results for the time's extraction form considered articles from Wikis. In this process three main categories taken that are news data [26], Warfare[26], and Celebrities[26] to build a data for training process summing up all articles yields a total of 1800 sentences.

To train purpose sampled 40 documents with 1600 sentences, within this test set there are total 268 events identified by Evita and total times are 148 hands coded manually the extracted times with the methods presented in table-3.

In Table-3 , the number of time expressions from a given input in first row and the second row explains retrieved times after executing each methods of the time extraction modules, the next row gives the relevant items from the retrieved, and last three rows are accuracy measures.

Table.3. Results obtained for the Methods

		Times Retrieved from the Input Data		
Test sources/ Articles	**No.of Events**	**Using SUTIME**	**Using our Pattern rules Algorithm**	**SUTIME + Holiday Package of our Algorithm**
Warfare	102	28	30	32
Celebrities	78	15	39	32
News data	88	39	69	61

Table.4 Accuracy of the Results (with Precision, Recall, F-measure)

	Using SUTIME	*SUTIME+ Holiday Package of our Algorithm*	*Using our Pattern rules Algorithm*
Total number of times in the given input	**148**	**148**	**148**
Total Retrieved times	**100**	**125**	**138**
Number of Relevant times from retrieved	**63**	**118**	**122**
Precision	**63%**	**94%**	**89%**
Recall	**43%**	**79%**	**82%**
F-measure	**51%**	**85.80%**	**85.30%**

Table-4 Shows obtained results for the comparing the three method's accuracies, using SuTime with holiday package obtained better precision that is 94% and 85% f-measure and our proposed algorithm for times is achieved equivalent f-measure with SuTime with holiday package results. With the above comparisons, the results of our approach also obtained noticeable precision.

6.3 Testing for Accuracy with Classifiers:

In the testing process, we feed the model with our test split documents. The model predicts the appropriate class of the model. Based on the predicted class and actual class of the document the confusion matrix is created. To measure the accuracy of the model uses metrics are 5fold cross-validation.

Table. 5. Comparison of accuracy parameters with different classifiers

Model Name	**Random Forest**	**Multinomial Naive Bayes**	**Logistic Regression**	**Linear SVC**
Accuracy	**53%**	**55%**	**68%**	**90%**
Precision	**31%**	**45%**	**65%**	**88%**
Recall	**27%**	**31%**	**45%**	**81%**
F-Score	**26%**	**33%**	**49%**	

6.5 Performance using various classifiers:

Random Forests perform slightly worse in the following cases:

- When the dimensionality (number of features) is very high with respect to the number of training samples.
- They fail in sharp corners and exactness. They use diffusion methods. They fit lumpy things well. They do not fit elaborate and highly detailed things well when the sample size is low.

Logistic Regression and Multinomial Naive Bayes

- These models are traditional models and they perform modestly better over Random Forests in NLP but very much far away from SVC's

Support Vector Machines: SVM outperforms all the other models by a large margin mainly because:

- With SVM regularization parameter, it avoids the problem of overfitting.
- Convex optimization problems resolved.
- Finally the approximation is to test error rate, in this case support vector machines gives the best accuracy in proposed model.

VII. CONCLUSION AND FUTURE SCOPE

We have successfully identified and classified events with 90% accuracy with effective pre-processing techniques and model. More amounts of data, the accuracy can be further improved. With 80 file got an accuracy of 85%. With 160 files we got an accuracy of 90%. The LinearSVC model performed better than all other models. In the observations of the work is the fact that a better collection of data helped us in training the model better.

Future Scope

Further if the data is more it may improve the accuracy. With more data, deep learning models which work very well on huge amounts of data. And this work can also make this as a web service which takes the article as an input and gives the event class and important dates of the article and identifies event time relationships. This model can be deployed in any real-time application to help readers to get articles based on the categories.

REFERENCES

1. Allan, J. 2002. Topic Detection and Tracking: Event-Based Information Organization, vol. 12.
2. Salton, G., Wong, A., and Yang, C. 1975. A vector space model for automatic indexing Communications of the ACM 18(11): 613–20.

3. Hatzivassiloglou, V., Gravano, L., and Maganti, A. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, pp. 224–31.
4. Makkonen, J., Ahonen-Myka, H., and Salmenkivi, M. 2004. Simple semantics in topic detection and tracking. Information Retrieval 7(3): 347–68.
5. Ahn, D., Schockaert, S., De Cock, M., and Kerre, E. 2006. Supporting temporal question answering: strategies for online data collection. In Proceedings of the 5th International Workshop on Inference in Computational Semantics. Buxton, UK: ACL, pp. 127–32.
6. ALLAN, J., R. PARKA, and LAVRENKO V. 1998. On-line new event detection and tracking. *In* Proceedings of the 21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval,SIGIR '98, ACM, New York, NY, pp. 37–45.
7. Grishman, R., and Sundheim, B. 1996. Message understanding conference-6: a brief history. In Proceedings of International Conference on Computational Linguistics (COLING 1996), Copenhagen, Denmark, vol. 96. Dresden, Germany: ICCL, pp. 466–71.
8. Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. 1998. University of She eld: description of the LaSIE-II system as used for MUC-7. In Proceedings of the Seventh Message Understanding Conferences (MUC-7), San Diego, CA. Gaithersburg, MD: NIST.
9. Aone, C., and Ramos-Santacruz, M. 2000. REES: a large-scale relation and event extraction system. In Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, WA. Stroudsburg, PA: ACL, pp. 76–83.
10. Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. 2000. Automatic acquisition of domain knowledge for information extraction. In Proceedings of the 18th Conference on Computational Linguistics, Hong Kong, vol. 2. Stroudsburg, PA: ACL, pp. 940–6.
11. Zhen Lei, Ying Zhang, Yu-chi Liu, et al. A system for detecting and tracking internet news event. In Advances in Multimedia Information Processing-PCM 2005, pages 754-764. Springer, 2005.
12. S. Bethard and J. H. Martin, "Identification of event mentions and their semantic class," presented at the Empirical Methods in Natural Language Processing (EMNLP), 2006.
13. R. Sauri, et al., "Evita: a robust event recognizer for QA systems," presented at the Human Language Technology and Empirical Methods in Natural Language Processing, 2005Saur´ı et al., 2005), Piskorski et al. (2007), REES (Aone and Ramos-Santacruz, 2000), Nishihara et al. (2009), Hung et al. (2010)
14. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project.In Proceedings of the COLING-ACL, Montreal. FrameNet (Baker et al. 1998) or VerbNet (Kipper et al. 2000), BioNLP (Jayalakshmi et al. 2014)
15. Chinchor, Nancy. 1999. Overview of MUC-7/MET-2. In Proc. Message Understanding Conference MUC-7.
16. Kirk Roberts,Bryan Rink, and Sanda M Harabagiu. A exible framework for recognizing events, temporal expressions, and temporal relations in clinical text. Journal of the American Medical Informatics Association, 20(5):867{875, 2013.TempEx Mani and Wilson,2000
17. Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In LREC, pages 3735{3740, 2012SUTime Changand Manning, 2012
18. Jannik Strotgen and Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 321-324,2010.
19. Pustejovsky, James, Jos´e M. Casta˜no, Robert Ingria,Roser Saur´?, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In IWCS-5.
20. J. Pustejovsky, et al., "TimeML: Robust Specication of Event and Temporal Expressions in Text." in New Directions in Question Answering, 2003J. Pustejovsky, et al., "The TIMEBANK corpus," 2003.Pustejovsky et al. 2003b.
21. Mani, I., B. Wellner, M. Verhagen, and J. Pustejovsky. 2007. Three Approaches to Learning TLINKs in TimeML. Technical Report CS-07-268, Computer Science Department, Brandeis University, Waltham,USA.
22. ACE. 2005. Evaluation of the Detection and Recognition of ACE: Entities, Values, Temporal Expressions, Relations, and Events. Gaithersburg, MD: NIST.
23. Verhagen, M.; Gaizauskas, R.; Schilder, F.; Hepple, M.;Katz, G.; and Pustejovsky, J. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In SemEval-2007: 4th International Workshop on Semantic Evaluations Verhagen et al. 2007).
24. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med InformAssoc. 2013;20(5):806–813. doi:10.1136/amiajnl-2013-001628.
25. TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text Naushad UzZaman Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 276–283.
26. en.wikipedia.org/wiki/Warfare:Warfare, Celebritie, and News data.
27. "Event Extraction using support vector machines" - FIRE 2017.
28. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G.,Pustejovsky, and J.: SemEval-2007 Task 15: TempEval Temporal Relation Identification. Proceedings of the 4th International Workshop on Semantic Evaluations (semEval-2007), pp. 75-80, Prague.
29. http://nlp.standford.edu/software/lex-parser.shtml
30. http://semeval2.fbk.eu/semeval2.php?location=download&task_id=5 &datatype=trial
31. Strotgen, J., Gertz, M.: Heideltime: high quality rule-based extraction and normalization oftemporal expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324. Association for Computational Linguistics (2010).

AUTHORS PROFILE

Mrs Vanitha Guda Assistant professor at Chaithanya Bharathi Institute Of Technology(A), Hyderabad. She is the Life member of Computer Society India- CSI, ISTE. Her Research interests are Natural Language Processing, Question Answering and published 15 papers in International, National journals and conferences.

Dr. S Suresh Kumar is presently working as Assistant professor and Head in the Department of Information Technology at JNTUH College of Engineering Jagtial. His area of specialization includes Natural Language Processing, Text summaries, etc. He published nearly 20 publications in various journals and conferences.