USING VOTING APPROACH FOR EVENT EXTRACTION AND EVENT-DCT , EVENT-TIME RELATION IDENTIFICATION

Anup Kumar Kolyal[1*] Asif Ekbal[2+] and Sivaji Bandyopadhyay[1#]

[1]Departmen of Computer Science and Engineering
Jadavpur University, India-700 032
*anup.kolya@gmail.com and #sivaji@cse.jdvu.ac.in
[2]Department of Computer Science and Engineering
Indian Institute of Technology Patna, India-800 013
+asif.@iitp.ac.in

ABSTRACT

Temporal information extraction is a popular and interesting research field in the area of Natural Language Processing (NLP) applications such as summarization, question answering (QA) and information extraction. In this paper, we have reported extraction of events and identification of different temporal relations between event-time and even-document creation time (DCT) within the TimeML framework. Our long term plan is to make temporal structure that can be used in the applications like question answering, textual entailment, summarization etc. In our approach, we propose a voted approach for (i) event extraction (ii) event – document creation time (DCT) relation identification (iii) event – time relation identification from the text under the TempEval-2 framework. The contributions of this work are two-fold; initially features are extracted from the training corpus and used to train a CRF and SVM framework. Then, the proposal of a voted approach for event extraction, event-DCT and event-time relation identification by combining the supervised classifiers such as Conditional Random Field (CRF) and Support Vector Machine (SVM). In total we generate 20 models, 10 each with CRF and SVM, by varying the available features and/or feature templates. All these 20 models are then combined together into a final system by defining appropriate voting scheme.

KEYWORDS

Temporal Relation Identification, Event, TimeML, Conditional Random Field, Support Vector Machine, TempEval-2010, WordNet.

1. INTRODUCTION

New sources of textual information, rich in events, grow significantly, such as social networks, blogs, and wikis. They are added to old sources like the informative web sites, emails and forums, which shows the importance to manage these data automatically. One of the important tasks of text analysis clearly requires identifying events, order of events described in a text and locating these in time. Event extraction has emerged to be very important in improving complex natural language processing (NLP) applications such as automatic and multidocument summarization [1] and question answering[2] (QA).Events are also described in different newspaper texts, stories and other important documents where events happen in time and ordering of these events are specified. In the context of summarization, identifying such events and relative order of events may help in getting better summaries .It is required to merge relative order of events so as any specific happenings have to be focused from multiple news sources correctly. In the QA system,

DOI : 10.5121/ijaia.2013.4106

to address questions when an event occurs, or what events occurred prior to a particular event [3].It relates not only to linguistics but also different scientific areas such as philosophy, psychology, etc. In NLP, different definitions of event can be found regarding the target application. On the one hand, in topic detection and tracking[4], event is defined as an instance of a topic identified at document level describing something that happen (e.g., "accident"). The aim of this task is to cluster documents on the same topic, that is to say, the same event. On the other hand, information extraction (IE) provides finer granularity event definitions. IE proposes standard schemes to annotate the individual events within the scope of a document. Sheffield Temporal Annotation scheme (STAG)[5] was aimed to identify events in news and their relationship with points in a temporal line. TimeML[6] presented a rich specification for annotating events in natural language (NL) text extending the features of the previous one.

Let us consider an example:

(1) *Prime Minister Benjamin Netanyahu* ***told*** *his Cabinet on* **Sunday** *that Israel was* ***willing*** *to* ***withdraw*** *from southern Lebanon.*

Here event identification task can identify three events which are '***told***', '**willing**' and '**withdraw**' and a temporal expression **'Sunday'** from example 1.Next, to identify the order of events in the time scale .We follow our event order identification task to quickly identify the ordering of events as: ***told→ willing→withdraw*** even though *relation types* '**overlap'** or '**before-or-overlap'** never appeared in the text.

We have followed TempEval-1 and TempEval-2 challenge attempts to address this question by establishing a common corpus on which research systems competed to find temporal relations [7]. TempEval-1 considers the following types of event-time temporal relations:

Task A: Relation between the events and times within the same sentence.

Task B: Relation between events and document creation times.

Task C. Relation between verb events in adjacent sentences.

TempEval-2 considers the following types for event-event and event-time temporal relation tasks:

A. Task was to determine the extent of the time expressions in a text as defined by the TimeML timex3 tag. In addition, values of the features type and val were to be determined. The possible values of type are time, date, duration, and set; the value of val is a normalized value as defined by the timex2 and timex3 standards

B. Task was to determine the extent of the events in a text as defined by the TimeML event tag.

C. Task was to determine the temporal relation between an event and a time expression in the same sentence.

D. Temporal relation between an event and the document creation time was to be determined.

E. Temporal relation between two main events in consecutive sentences.

F. Temporal relation between two events, where one event syntactically dominates the other event.

In the above each temporal relation identification task of TempEval-1 and TempEval-2(Task A and Task B are not included), systems attempt to annotate appropriate pairs with one of the following relations: BEFORE, BEFORE-OR-OVERLAP, OVERLAP, OVERLAP-OR-AFTER, AFTER or VAGUE. The participating teams were instructed to find all temporal relations of these types in a corpus of newswire documents. It is described at [8].

In this work, we propose a voted approach for (i) event extraction (ii) event – document creation time (DCT) relation identification (iii) event – time relation identification from the text under the TempEval-2 framework. The contributions of this work are two-fold, namely (i) for event extraction the proper identification and use of morphological, syntactic, and lexical semantic features; on the other hand for event-DCT and event-time relation identification pairs are encoded using syntactically and semantically motivated features present in the TempEval-2 corpus and identified context feature from corpus. These features have been automatically extracted from the training corpus and used to train a CRF and SVM framework. It is to be noted that we have only used some of the features available in the training corpus. (ii) the proposal of a voted approach for event extraction, event-DCT and event-time relation identification by combining the supervised classifiers such as Conditional Random Field (CRF) and Support Vector Machine (SVM). Initially, we developed event extraction,event-dct and event-time systems using these supervised classifiers. We analyzed the performance of each of these systems by considering the various available feature combinations. These features are mostly extracted from the gold standard TempEval-2 corpus. Thereafter, we identify various useful features from the semantic resource like WordNet, semantic roles ,a number of heuristics that are defined based on the inflection information of the word tokens and depending on the various context features. In total we generate 20 models, 10 each with CRF and SVM, by varying the available features and/or feature templates. All these 20 models are then combined together into a final system by defining appropriate voting scheme. Evaluation results with the TempEval-2 evaluation challenge [9] for event extraction yield the precision, recall and F-measure values of 86.10%, 84.90% and 85.50%, respectively. This is actually an improvement of 3.5 percentage F-measure points over the best performing system[10]of TempEval-2 and event-DCT[11],event-time[12].

The rest of the paper is organized as follows. Section 2 presents the related works. Section 3 very briefly discussed about TimeBank Corpus Section 4 described approach for Event Extraction , Event-DCT and Event-Time relation identification Section 5 Features used for Event Extraction Section 6 Features used for Event DCT Relation Identification Section 7 Features used for Event Time Relation Identification Section 8 Evaluation Results for Event Extraction Section 9 Evaluation Results for Event DCT Relation Identification Section 10 Evaluation Results for Event time Relation Identification Section 11 Voting Techniques Section 12 Conclusion.

2. RELATED WORKS

It is worth noting that the event definition varies according to the application domain: probabilities, software development, history, philosophy and linguistics. But it can be said that an event is something that happens, it can frequently be described as a change of state or a transition. Automatic Content Extraction (ACE) definition adds that an event is a specific occurrence involving participants. TimeML specification [13] considers 'event' as a cover term for situations that happen or occurs. Events can be punctual or last for a period of time. TimeML also considers as events those predicates describing states or circumstances in which something obtains or holds true. The tasks of event extraction were first explored in the series of Message Understanding Conferences (MUCs) started from 1987. The events in MUCs were limited to finite topics, e.g., terrorist activities, management succession. The existing works on event extraction are based either on pattern-matching rules [14], or on the machine learning approach [15]. Different systems represent events in different ways. The existing approaches are TimeML and ACE models. In TimeML, an event is a word that points to a node in a network of temporal relations.

[16] Described a system to identify events and semantic class of these events mainly for question answering purpose. They introduced several linguistic features like text, affix, morphological, word class, negation, wordnet hypernym and surrounding word or phrase information features to train SVM. Their system STEP (System for Textual Event Parsing) achieved 82.0% precision almost 71% recall.

Like event extraction, a wealth of prior research has been done for temporal relation identification. Various machine learning algorithms has been applied to formulate the temporal relation as an event paired with a time or another event and translated these into a set of feature values. Some of the popularly used machines learning techniques were Naive-Bayes, Decision Tree (C5.0), Maximum Entropy (ME) and Support Vector Machine (SVM). Machine learning techniques alone cannot always yield good accuracies. To achieve reasonable accuracy, some researchers [17] used hybrid approach. The basic principle of hybrid approach is to combine the rule-based component with machine learning. It has been shown in [17] that classifiers make most mistakes near the decision plane in feature space. The authors carried out a series of experiments for each of the three tasks on four models, namely naive-Bayes, decision tree (C5.0), maximum entropy and support vector machine. The system was designed in such a way that they can take the advantage of rule-based as well as machine learning during final decision making. But, they did not explain exactly in what situations machine learning or rule based system should be used given a particular instance. They had the option to call either component on the fly in different situations so that they can take advantage of the two empirical approaches in an integrated way.

[18] Reported an automatic event identification system based on SVM. It obtained the F-measure value of 76.4% in a 10-fold cross validation experiment on the Time-Bank corpus.

[10] Reported a CRF based system for event recognition. They achieved 81.4% F-score with the features based on morphosyntax, ontology and semantic roles.

In [7], CU-TMP participant used gold-standard TimeBank features and along with this for task A time related preposition governing syntactic features , for task B auxiliaries governing event and event’s stem features , for task C verb and auxiliaries governing the second event syntactic features which are derived from text to train three SVM model. LCC-TE participant used (i) syntactic pattern matching tool based on hand-crafted finite state rule for temporal expression identification and normalization(ii)set of heuristics rule based on lexicon, lemma, part of speech and wordnet senses for event detection(iii)for temporal relation identification large set of syntactic and semantic features as input to a machine learning components. NAIST-Japan participant used dependency trees features which are extracted from a dependency parser for temporal relation identification.

3. TIMEBANK CORPUS

In this section we have briefly described about TimeBank corpus which one has been taken as platform of doing work. This is the only TimeML reference corpus of annotated news articles. It is a rich research resource of different kind complex temporal, event expression and their temporal relationship. This is one of the planned efforts to develop from a systematic, linguistically grounded approach to an annotation-based framework for analysis of time in text [19]. The corpus has two version of release. First version of release was developed within the TERQAS initiative and motivation behind the second version of release is to answer temporal based question about the events, which enhance the natural language question answering systems.

Initially, TERQAS was introduced for the computational analysis of time, basically various kind of time stamping and temporal ordering of events and/or relations within a narrative that is required for information techniques. These challenges comes from text to a rich representation of temporal entities, ontologically grounded temporal graph and reasoning capability

The main goal of TERQAS was to represent a framework for distinguishing events and their temporal ordering in text. Actually this framework helped to make temporal analysis algorithms. This analytical knowledge was common base for markup language of time. Then development, testing and evaluation annotated corpus set are driven by TERQAS defined TimeML language.

By the number of research efforts TimeBank 1.1[20] corpus has been effectively leveraged and annotated with the TimeML complaint parsing. Temporal-event annotated corpus TIMEBANK contains [186] news articles which is a collection of temporal-event annotated corpus with annotations of terms denoting events, temporal expressions, and temporal signals, and, most importantly, of links between them denoting temporal relations. The TimeBank 1.1 is the first stable version of sizable experimental training corpus for ML-based language algorithms. It is both exercised and stressed with the expressive equipment of TimeML 1.1. More cleaner and consistent corpus based on TimeML 1.2 is TimeBank 1.2 which is also robust and more expressive.

Whatever, there is, a significant difference in status between TimeML and TimeBank1.1. (respectively [6] and [21]) TimeBank 1.2, beside that, it is almost a ‘side effect’ of the TERQAS work.

In TIMEX3 more details analysis about event class classification and some discourse level TLINKs analysis than, TIMEX2 describe at [15] [22][23].

It is not straightforward extension of TIMEX3 from TIMEX2.TIMEX2 and TIMEX3 vary to a large extent in their behavior of event anchoring and set of times. Specially, relational time expressions (e.g., 5 days after departure) are a single in the TIMEX2 format; under TimeML analysis, the same expression would be annotated as a group of related TIMEX3, SIGNAL and EVENT tags,with an additional LINK anchoring the EVENT.

4. Our approach for event extraction and event-DCT, Event-time relation identification

In this work, we report our works on (i) event extraction (ii) event – document creation time (DCT) relation identification (iii) event – time relation identification using a voted technique. Initially, a number of different models are generated using Conditional Random Field (CRF) and Support Vector Machine (SVM). All these models are then combined together into single system by defining appropriate combination techniques. We use two voting techniques: (i). majority voting where same weight is assigned to all the component classifiers and (ii). weighted voting where the individual classifiers are assigned weights according to their performance. Brief descriptions of the base classifiers, i.e. CRF and SVM are presented in the following subsections

4.1. Conditional Random Field

Conditional Random Field (CRF) [24] is an undirected graphical model, which is a special case of which corresponds to conditionally trained probabilistic finite state automata. The main advantage of CRF comes from that it can relax the assumption of conditional independence of the observed data often used in generative approaches, an assumption that might be too restrictive for a considerable number of object classes. Additionally, CRF avoids the label bias problem.

CRF is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $S = < s_1, s_2, ..., s_T >$ given an observation sequence $O = < o_1, o_2, ..., o_T)$ is calculated as:

$$P_\Lambda(s \mid o) = \frac{1}{Z_o} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)\right)$$

where, $f_k(s_{t-1}, s_t, o, t)$ is a feature function whose weight λ_k is to be learned via training. The values of the feature functions may range between $-\propto .....+\propto$, but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor,

$$Z_0 = \sum_s \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)\right),$$

This, as in HMMs, can be obtained efficiently by dynamic programming. Here, the CRF parameters are optimized using Limited-memory BFGS [25], a quasi-Newton method that is significantly more efficient, and results in only minor changes in accuracy due to changes in σ. CRFs generally can use real-valued functions but it is often required to incorporate the binary valued features. A feature function $f_k(s_{t-1}, s_t, o, t)$ has a value of 0 for most cases and is only set to 1, when s_{t-1}, s_t are certain states and the observation has certain properties. We use the C^{++} based CRF++ package[1], a simple, customizable, and open source implementation of CRF for segmenting /labeling sequential data.

4.2. Support Vector Machine

Support Vector Machines (SVMs)[26] are relatively new machine learning approaches for solving two-class pattern recognition problems. SVMs are well known for their good generalization performance, and have been applied to many pattern recognition problems. In the field of NLP, SVMs are applied to text categorization, and are reported to have achieved high accuracy without falling into over-fitting even though with a large number of words taken as the features.

Suppose we have a set of training data for a two-class problem: $\{(x_1, y_1), ...(x_N, y_N)\}$, where $x_i \in R^D$ is a feature vector of the i-th sample in the training data and $y_i \in \{+1, -1\}$ is the class to which x_i belongs. The goal is to find a decision function that accurately predicts class y for an input vector x. A non-linear SVM classifier gives a decision function $\mathrm{f(x) = sign(g(x)}$ for an input vector where,

$$g(x) = \sum_{i=1}^{m} w_i K(x, z_i) + b$$

Here, $\mathrm{f(x)} =$ +1 means x is a member of a certain class and $\mathrm{f(x)} =$ -1 means x is not a member. z_i s are called support vectors and are representatives of training examples, m is the number of support vectors. Therefore, the computational complexity of $g(x)$ is proportional to m. Support vectors and other constants are determined by solving a certain quadratic programming problem. $K(x, z_i)$ is a *kernel* that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products: $K(x, z_i) = k(x.z)$.

[1]http://crfpp.sourceforge.net

A polynomial kernel of degree d is given by $K(x, z_i) = (1+x)^d$. We can use various kernels, and the design of an appropriate kernel for a particular application is an important research issue.

We have developed our system using SVM [27] and [26], which performs classification by constructing an N-dimensional hyperplane that optimally separates data into two categories. Our general NER system includes two main phases: training and classification. Both the training and classification processes were carried out by *YamCha*[2] toolkit, an SVM based tool for detecting classes in documents and formulating the event extraction task as a sequential labeling problem. We use the *polynomial kernel function* and TinySVM-0.07[3] classifier.

5. FEATURES USED FOR EVENT EXTRACTION (EE)

The individual classifiers are trained and tested with the features that range from morphological, syntactic and to lexical semantic. All these features are extracted from the TempEval gold standard datasets, semantic knowledge base like WordNet, semantic role labels and from the various heuristics that are defined based on the nature of the available corpus.

5.1. Syntactic Features

These features are extracted from the gold-standard TimeBank corpus. In the present work, we mainly use the various combinations of the following features:

(i). **Part of Speech (POS) of event terms**: It denotes the POS information of the event. The features values may be either of ADJECTIVE, NOUN, VERB, and PREP.
(ii). **Event Tense**: This feature is useful to capture the standard distinctions among the grammatical categories of verbal phrases. The tense attribute can have values, PRESENT, PAST, FUTURE, INFINITIVE, PRESPART, PASTPART, or NONE.
(iii). **Event Aspect**: It denotes the aspect of the events. The aspect attribute may take values**,** PROGRESSIVE, PERFECTIVE and PERFECTIVE PROGRESSIVE or NONE.
(iv). **Event Polarity**: The polarity of an event instance is a required feature represented by the boolean feature, polarity. If it is set to 'NEG', the event instance is negated. If it is set to 'POS' or not present in the annotation, the event instance is not negated.
(v). **Event Modality**: The modality feature is only present if there is a modal word that modifies the instance.
(vi). **Event Class**: This is denoted by the 'EVENT' tag and used to annotate those elements in a text that mark the semantic events described by it. Typically, events are verbs but can be nominal also. It may belong to one of the following classes:
REPORTING: Describes the action of a person or an organization declaring something, narrating an event, informing about an event, etc.
PERCEPTION: Includes events involving the physical perception of another event. Such events are typically expressed by verbs like: *see*, *watch*, *glimpse*, *behold*, *view*, *hear*, *listen*, *overhear* etc.
ASPECTUAL: Focuses on different facets of event history.
I_ACTION: An intentional action. It introduces an event argument which must be in the text explicitly describing an action or situation from which we can infer something given its relation with the I_ ACTION.
I_STATE: Similar to the I_ACTION class. This class includes states that refer to alternative or possible words, which can be introduced by subordinated clauses, nominalizations, or untensed verb phrases (VPs).
STATE: Describes circumstances in which something obtains or holds true.

[2]http://chasen-org/~taku/software/yamcha/
[3]http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM

Occurrence: Includes all of the many other kinds of events that describe something that happens or occurs in the world.

(viii). **Event Stem**: It denotes the stem of the head event.

5.2. WordNet Features

WordNet [28] features have been widely 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. Here, we use WordNet mainly to identify *non-deverbal event nouns* like '*war*', '*attempt*', '*tour*' etc. These words have noun (NN) PoS information. We know from the lexical information of WordNet that the words like '*war*' and '*tour*' are generally used as both *noun* and *verb* forms in the sentence. We design two following rules based on the WordNet:

Rule 1: The word tokens having Noun (NN) PoS categories are looked into the WordNet. If it appears in the WordNet with noun and verb senses, then that word token is also considered as effective feature of event identification. For example, *war* has both noun and verb senses in the WordNet, and thus considered as an event feature.

Rule 2: The *stems* of the noun word tokens are looked into WordNet. If one of the WordNet senses is verb then the token will be identified as verb. For example, the stem of *proposal*, i.e. *propose* has two different senses, noun and verb in the WordNet, and thus it is considered as an event feature.

We observe significant performance improvement on event extraction with the above mentioned two rules and using these as features of the supervised classifiers, CRF and SVM.

5.3. Inflection Features

We used WordNet to extract the feature of event expressions that appear in the WordNet with both noun and verb senses. Here, we mainly concentrate to identify the specific lexical classes like '*inspection*' and '*resignation'*. These can be identified by the suffixes such as ('*-ción*'), ('-*tion*') or ('*-ion*'), i.e. the morphological markers of deverbal derivations.

Initially, we run the CRF based Stanford Named Entity (NE) tagger[4] on the TempEval-2 test dataset. The output of the system is tagged with *Person*, *Location*, *Organization* and *Other* classes. The words starting with the capital letters are also considered as NEs. Thereafter, we came up with the following rules for feature extraction of event:

Cue-1: Nouns which are morphologically derived from verbs are commonly distinguished as nominalizations (or, deverbal nouns). The deverbal nouns are usually identified by the suffixes like '*-tion*', '*-ion*', '*-ing*' and '*-ed*' etc. The nouns that are not NEs, but end with these suffixes are considered as the event feature.

Cue 2: The verb-noun combinations are searched in the sentences of the test set. The non-NE noun word tokens are considered as the event features.

Cue 3: Nominals and non-deverbal event nouns 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 word token(s) appearing after these clue word(s)/phrase(s) are considered as events feature.

Cue 4: The non-NE nouns occurring after the expressions such as *frequency of*, *occurrence of* and *period of* are most probably the event nouns feature.

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Cue 5: Event nouns can also appear as objects of aspectual and time-related verbs, such as *have begun a campaign* or *have carried out a campaign* etc. The non-NEs that appear after the expressions like "*have begun a*", "*have carried out a"* etc. are also considered as event feature.

5.4. Features using Semantic Roles

We use Semantic Role Label (SRL) [29][30][31] to identify different features of the sentences of a document. These features help us to extract the events from the text. For each predicate in a sentence acting as event word, semantic roles extract all constituents, determining their arguments (agent, patient, etc.) and their adjuncts (locative, temporal, etc.). Some of the other features like predicate, voice and verb sub-categorization are shared by all the nodes in the tree. In the present work, we use predicate as an event. Semantic roles can be used to detect the events that are the nominalizations of verbs such as *agreement* for *agree* or *construction* for *construct*. Event nominalizations often afford the same semantic roles as verbs, and often replace them in written language [32]. Nominalisations (or, *deverbal nouns*) are commonly defined as nouns, morphologically derived from verbs, usually by suffixation [33]. They can be classified into at least three categories in the linguistic literature, event, result, and agent/patient nominalisations [34]. Event and result nominalisations constitute the bulk of *deverbal* nouns. The first class refers to an event/activity/process, with the nominal expressing this action (e.g., killing, destruction etc.). Nouns in the second class describe the result or goal of an action (e.g., agreement, consensus etc.). Many nominals have both an event and a result reading (e.g., selection). A smaller class is agent/patient nominalizations that are usually identified by suffixes such as *-er*, *-or* etc., while patient nominalisations end with *-ee*, *-ed* (e.g. employee). Let us consider the following example sentence to understand how semantic roles can be used for event extraction.

All sites were inspected to the satisfaction of the inspection team and with full cooperation of Iraqi authorities, Dacey said.

The output of SRL for this sentence is as follows:

[ARG1 All sites] were [TARGET inspected] to the satisfaction of the inspection team and with full cooperation of Iraqi authorities, [ARG0 Dacey] [TARGET said]

The sentence is traversed to find the argument-target relations. A sentence is scanned as many times as the number of target words in the sentence. In the first traversal, *inspected* is identified as the event. In the second pass, *said* is identified as an event. All the extracted target words are treated as the event words. We observed that many of these target words are identified as the event expressions by the CRF and SVM models. But, there exists many nominalised event expressions (i.e., *deverbal nouns*) that are not identified as events by the supervised CRF and/or SVM. These nominalised expressions are correctly identified as events by SRL. We effectively use this semantic role information as a feature of the classifier, and observe significant performance improvement.

6. Features used for Event DCT Relation Identification

We have used the gold-standard TimeBank features for events and times for training the CRF and SVM

6.1. Available Features

In the present work, we mainly use the various combinations of the following features:

(i) Part of Speech (POS) of event terms: It denotes the POS information of the event. The features values may be either of ADJECTIVE, NOUN, VERB, and PREP.

(ii) Event Tense: This feature is useful to capture the standard distinctions among the grammatical categories of verbal phrases. The tense attribute can have values, PRESENT, PAST, FUTURE, INFINITIVE, PRESPART, PASTPART, or NONE.
(iii) Event Aspect: It denotes the aspect of the events. The aspect attribute may take values**,** PROGRESSIVE, PERFECTIVE and PERFECTIVE PROGRESSIVE or NONE.
(iv)Temporal Relation between the DCT and the Temporal Expression in the target sentence: The value of this feature could be “greater than”, “less than”, “equal”, or “none”.

6.2. Derived Features

We have identified different types of context based syntactic features which are derived from text to distinguish the different types of temporal relations. In this task, following features help us to identify the events and DCT specially “AFTER” temporal relation:

(i) Modal Context: Whether or not the event word has one of the context words like, *will*, *shall*, *can*, *may*, or any of their variants (might, *could*, *would*, etc.).In this sentence: “The entire world **will** [EVENT ***see*]** images of the Pope in Cuba”. Here “will” context word helps us to determine event -DCT relation ‘AFTER’.
(ii) Preposition Context: Any prepositions before the event or time, we consider one example:”Children and invalids would be permitted **to** [EVENT leave**]** **Iraq”**. Here the preposition *to* help us to determine event-DCT relation ‘AFTER’. In the same way for time also: *on Friday* and *for nearly forty years*, the prepositions *on* and *for* governs the time.
(iii) Context word before or after temporal expression: context word like before, after, less than, greater than help us to determine event-time temporal relation identification .We considers one example:”**After** ten years of [EVENT boom**] ….”**

7. Features used for Event Time Relation Identification

We use the gold-standard TimeBank features for events and times for training the CRF and SVM. The features are listed below:

7.1. Available Features

In the present work, we mainly use the various combinations of the following features:
(i).**Event class**: Denoted by the ‘EVENT’ tag and used to annotate those elements in a text that mark the semantic events described by it.
(ii).**Event stem**: Denotes the stem of the head event.
(iii).**Event and time strings**: Denotes the actual event strings and time.
(iv).**Part of Speech of event terms**: Denotes the POS information of the event (e.g., ADJECTIVE, NOUN, VERB, PREP).
(v).**Event tense**: Captures standard distinctions among the grammatical categories of verbal phrases.
(vi).**Event aspect**: Denotes the aspect attribute of the event that may take values**,** PROGRESSIVE, PERFECTIVE and PERFECTIVE PROGRESSIVE or NONE.
(vii).**Event polarity**: Polarity of an event instance is represented by the boolean value, POSITIVE or NEGATIVE.
(viii).**Event modality**: The modality attribute is only present if there is a modal word that modifies the instance.

7.2. Derived Features

Like event-DCT relation type, here we have also identified different types of context based temporal expression features which are derived from text to distinguish the different types of temporal relations. In this task, following features help us to identify between events and time specially "AFTER" and "BEFORE" temporal relation. Following features are derived from text.

(i) **Type of temporal expression**: Represents the temporal relationship holding between events, times, or between an event and a time of the event.
(ii).**Temporal signal**: Represents temporal prepositions "*on*" (*on this coming Sunday*) and slightly contribute to the overall score of classifiers
(iii).**Temporal Expression in the target sentence**: Takes the values *greater than*, *less than*, *equal* or *none*. These values contribute to the overall score of classifiers.

8. Evaluation Results for Event Extraction

We use the TempEval-2 datasets to report the evaluation results. Initially, a number of various models of CRF and SVM are generated by varying the different features and/or feature templates. Finally, we have a training data in the form (W_i, T_i), where, W_i is the i^{th} pair along with its feature vector and T_i is its corresponding output label (i.e., *Event* or *Other*). Models are built based on the training data and the feature template.

Each of these classifiers is evaluated with the TempEval-2 gold standard test dataset. The test data had 373 verbal and 125 non-deverbal event nouns. Evaluation results for CRF are shown in Table 1. Though we constructed 10 different classifiers by varying the available features, the performance of the best CRF model is shown in the table. It shows the precision, recall and F-measure values of 75.30%, 78.10% and 76.67%, respectively. We observe this highest F-measure with the context of size five, i.e. preceding two and following two words, bigram of current and previous token, bigram of current and next token, dynamic event information of the previous token and the feature vector consisting of all the features of the current token only. Thereafter, we include the features extracted from WordNet, semantic roles and several heuristics. Table 1 shows how these features help to improve the overall performance. The use of semantic roles shows an increment of 2.11 percentage F-measure points over the system that uses only the gold standard TimeBank features. The WordNet feature is most effective and its use improves the performance by 3.49 percentage F-measure points. Finally, the CRF-based system achieves the performance with precision, recall and F-measure values of 82.90%, 84.20% and 83.54%, respectively.
Evaluation results for SVM based event extraction system are reported in Table 2. Out of 10 classifiers, we show the evaluation figure of only the best SVM model. It shows the precision, recall and F-measure values of 76.30%, 75.10% and 75.70%, respectively. The use semantic roles, WordNet and heuristics improve the F-measure values by 1, 5.9 and 1.34 percentage points, respectively. Results of both CRF and SVM suggest that WordNet is most effective to improve the overall performance of the system.

Table 1. Evaluation results of CRF-based event extraction (we report percentages

Model	precision	Recall	F-measure
CRF	75.30	78.10	76.67
CRF+SRL	77.60	80.00	78.78
CRF+SRL+WordNet	81.56	83.00	82.27
CRF + SRL + WordNet + Rules	82.90	84.20	83.54

Table 2. Evaluation results of SVM-based event extraction (we report percentages)

Model	precision	Recall	F-measure
SVM	76.30	75.10	75.70
SVM+SRL	77.30	76.10	76.70
SVM+SRL+WordNet	81.20	84.05	82.60
SVM + SRL + WordNet + Rules	84.70	83.20	83.94

9. EVALUATION RESULTS FOR EVENT DCT RELATION IDENTIFICATION

We develop a number of models of CRF and SVM based on the features included into it. A feature vector consisting of the available features as described in Section 5.1 and context word based syntactic features which are described in Section 5.2 to distinguish the different types of temporal relations are extracted for each <event, DCT> pair in the TimeBank corpus. Now, we have a training data in the form (W_i, T_i), where, W_i is the i^{th} pair along with its feature vector and T_i is it's corresponding TempEval relation class. Models are built based on the training data and the feature template.

Each of these classifiers is evaluated with the TempEval-2 gold standard test dataset. The test data had 190 event-DCT relation links within that 111 BEFORE, 24 AFTER, 45 OVERLAP, 4 BEFORE-OR-OVERLAP, 2 OVERLAP-OR-AFTER and 3 NON-RELATED Link. Evaluation results for CRF and SVM are shown in Table 3 and Table 4 respectively. Though we constructed 10 different classifiers by varying the available features, the performance of the best CRF and SVM model is shown in the table 5. An even simpler evaluation metric similar to the definition of 'accuracy' is used to evaluate for relation types event-DCT. The metric (henceforth referred to as 'accuracy') is defined as below: the number of correct answers divided by the number of answers.It shows accuracy of the above relation type 83.6%. We observe this highest accuracy for the feature vector consisting of current token and POS; combination of POS and tense of the current token, combination of polarity and POS of the current word, combination of POS and aspect of current word, combination of polarity and POS of current word, combination of POS, tense and aspect of the current token, combination of derived context word, pos, tense and current token.

During evaluation, we obtain the highest performance for the following feature template as shown in Figure 1. The test corpus of event-DCT relation type consists of 190 relational links from TimeBank[20]. The performance is assessed with only accuracy evaluation metrics.

Figure 1: Best Feature Template of the CRF and SVM based System

$W_{(i-2)}$
$w_{(i-1)}$
w_i
w_{i+1}
$w_{(i+2)}$
Combination of w_{i-1} and w_i
Combination of w_i and w_{i+1}
Dynamic output tag (t_i) of the previous pair
Feature vector of w_i of other features

Table 3

Feature Combination	Result
Context, Tense, POS	0.72
Context, Aspect, POS	0.65
Context, Tense, Aspect	0.82
Context, Tense, Aspect, Syntactic feature	0.836

Table 4

Feature Combination	Result
Context, Tense, POS	0.71
Context, Aspect, POS	0.67
Context, Tense, Aspect	0.814
Context, Tense, Aspect ,Syntactic feature	0.829

Table 3 and Table 4 are evaluation result of CRF and SVM using different feature combinations respectively

Evaluation results with different feature representations are reported in Table 3 and Table 4 for CRF and SVM respectively. Results show that the system performs better with the context of size five (i.e., previous two, current and the next two <event, DCT> pairs), tense, aspect and context word features. It shows highest accuracy 0.836.

Before relation types links are baseline model

Table 5: Evaluation results

Technique	Strict
	P
Baseline	0.584
CRF	0.836
SVM	0.829

The overall evaluation results of the system are presented in Table 5. It shows the results of the *baseline* model, CRF based system as well as the SVM-based system. The *baseline* model is developed based on the most frequent temporal relation encountered in the training data for the task. In the case of event-DCT the most frequent temporal relation present in the training data is BEFORE. Results show that the CRF and SVM based system performs better than the *baseline* model with the margins of 25.6% and 24.9% accuracy respectively. The CRF-based system performs best among all the models.

10. EVALUATION RESULTS FOR EVENT TIME RELATION IDENTIFICATION

We have included the features to develop various CRF and SVM models . A feature vector has been formed by the available features as described in Section 6.1 and derived features from 6.2 are extracted for each <event, time> pair in the TempEval-2 corpus. Now, we have a training data in the form (W_i, T_i), where, W_i is the i^{th} pair along with its feature vector and T_i is it's corresponding TempEval-2 relation class. Models are built based on the training data and the feature template.

During event-time relation identification, to obtain the highest performace of the best feature template for CRF and SVM model, we have followed figure 1. Template in section 8.

The test data had 65 event-time relation links within that 8 BEFORE, 10 AFTER, 41 OVERLAP, 1 BEFORE-OR-OVERLAP, 2 OVERLAP-OR-AFTER and 3 NON-RELATED Link. Evaluation results for CRF and SVM model are shown in Table 7 and Table 6 respectively. Though we constructed 10 different classifiers by varying the available features, the performance of the best CRF and SVM models are shown in the Table 8. An even simpler evaluation metric similar to the definition of ‘accuracy’ is used to evaluate for relation types event-time. The metric (henceforth referred to as ‘accuracy’) is defined as below: the number of correct answers divided by the number of answers.It shows accuracy of the above relation type 64.9%. We observe this highest accuracy for the feature vector consisting of current token and POS; combination of POS and tense of the current token, combination of polarity and POS of the current word, combination of POS and aspect of current word, combination of polarity and POS of current word, combination of POS, tense and aspect of the current token, type of temporal expression, temporal signal and temporal Expression in the target sentence.

During evaluation, we obtain the highest performance for the following feature template as shown in Figure 1. The test corpus of event-time relation type consists of 65 relational links from TimeBank [20]. The performance is assessed with only accuracy evaluation metrics.

Table 6

Feature Combination	Result
Context, Tense, POS	0.619
Context, Aspect, POS	0.591
Context, Tense, Aspect	0.63
Context, Tense, Aspect,combination of derived feature	0.649

Table 7

Feature Combination	Result
Context, Tense, POS	0.601
Context, Aspect, POS	0.573
Context, Tense, Aspect	0.619
Context,Tense, Aspect, combination of derived feature	0.638

Table 6 and Table 7 are evaluation result of SVM and CRF using different feature combinations Evaluation results with different feature representations are reported in Table 8 for CRF and SVM. Results show that the system performs better with the context of size five (i.e., previous two, current and the next two <event, time> pairs), tense, aspect, type of temporal expression, temporal preposition and temporal Expression in the target sentence features. Before relation types links are baseline model

Table 8: Evaluation results

Technique	Strict
	P
Baseline	0.630
CRF	0.649
SVM	0.638

The overall evaluation results of the system are presented in Table 8. It shows the results of the *baseline* model, CRF based system as well as the SVM-based system. The *baseline* model is developed based on the most frequent temporal relation encountered in the training data for the task. In the case of event-time the most frequent temporal relation present in the training data is OVERLAP. Results show that the CRF and SVM based system performs better than the *baseline* model with the margins of 1.9% and 0.8% accuracy respectively. The CRF-based system performs best among all the models.

11. VOTING TECHNIQUES

In order to obtain higher performance, we define appropriate mechanisms to combine several classifiers. All the CRF and SVM based classifiers are combined together into a final system by weighted voting. Each of these classifiers is built by varying the features and/or feature templates included into it. We define following two weighting methods:

(1). **Uniform weights (Majority voting):** All the models are assigned the same voting weight. The combined system selects the classifications, which are proposed by the majority of the models.

(2). **F-measure value**: Here, the F-measure value of the individual classifier is used as the weight of the corresponding classifier.

11.1. Voting Result for Event Extraction

Here we combine the classifiers only after including the features extracted from the gold standard corpus, semantic roles, WordNet and several heuristics.

Experimental results of the voted system are presented in Table 9. Evaluation results show that the system achieves the highest performance for the weighted voting scheme that considers F-measure as the weight of the classifier. Voting shows an overall improvement of **1.96%** over the CRF-based model and **1.56%** over the SVM-based model.

Table 9.Evaluation results of the voted system (we report percentages)

Model	precision	recall	F-measure
CRF [Best]	82.90	84.20	83.54
SVM[Best]	84.70	83.20	83.94
Majority voted	85.20	84.10	84.65
Weighted voted	86.10	84.90	85.50

11.2. Voting Result for Event-Dct relation identification

Here we combine the classifiers only after including the features extracted from the gold standard corpus and using some derived features like modal context and propositional context features.

Experimental results of the voted system are presented in Table 10. Evaluation results show that the system achieves the highest performance for the weighted voting scheme that considers F-measure as the weight of the classifier. Voting shows an overall improvement of **1.30%** over the CRF-based model and **2.00%** over the SVM-based model.

Table 10. Evaluation results of the voted system (we report percentages)

Model	precision	recall	F-measure
CRF [Best]	83.60	83.60	83.60
SVM[Best]	82.90	82.90	82.90
Majority voted	84.10	84.10	84.10
Weighted voted	84.60	84.90	84.90

11.3. Voting Result for Event-time relation identification

Here we also combine the classifiers only after including the features extracted from the gold standard corpus and using some derived features like type of temporal expression and temporal signal.

Experimental results of the voted system are presented in Table 11. Evaluation results show that the system achieves the highest performance for the weighted voting scheme that considers F-measure as the weight of the classifier. Voting shows an overall improvement of **1%** over the CRF-based model and **2.10%** over the SVM-based model.

Table 11. Evaluation results of the voted system (we report percentages)

Model	precision	recall	F-measure
CRF [Best]	64.90	64.90	64.90
SVM[Best]	63.80	63.80	63.80
Majority voted	65.40	65.40	65.40
Weighted voted	65.90	65.90	65.90

12. CONCLUSIONS

In this paper, we have reported our work on event extraction under the TempEval -2010 evaluation exercise. We proposed a voted approach for event extraction. A number of models based on CRF and SVM were generated by varying the available features and/or feature templates. These CRF and SVM based systems suffer mostly in identifying the deverbal nouns that denote the event expressions. Thereafter, we came up with several proposals in order to improve the system performance. We extracted many useful features from SRL, WordNet and handcrafted rules. Evaluation showed that all these features are very effective to improve the performance of each of the supervised classifiers. Finally, we combined all the individual classifiers by defining appropriate weighted voting techniques. Evaluation results yield the precision, recall and F-measure values of 86.10%, 84.90% and 85.50%, respectively. This is an improvement of approximately 3.00 percentage F-measure points over the best performing system of TemEval-2010 evaluation challenge.

Future works include the identification of more precise rules for event identification and multiword events. Future works also include experimentations with other machine learning techniques like maximum entropy and genetic algorithm.

ACKNOWLEDGEMENTS

The work is partially supported by a grant from English to Indian language Machine Translation(EILMT) funded by Department of Information and Technology (DIT), Government of India.

REFERENCES

[1] Daniel, Naomi, Dragomir Radev, and Timothy Allison.2003. *Sub-event based multi-document summariza-tion.* In HLT-NAACL Text summarization workshop, pages 9–16. ACL.

[2] Pustejovsky, James. 2002. TERQAS: *Time and EventRecognition for Question Answering Systems.* InARDA Workshop.

[3] Mani, I.,Wellner,B.,Verhagen,M.,Lee C.M.,Pustejovsky,J.,:Machine learning of temporal relation. In: Proceedings of the 44th Annual meeting of the Association for Computational Linguistics, Sydney, Australia, ACL,2006.

[4] Allan, James. 2002. Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publishers, Norwell, MA, USA.

[5] Setzer, Andrea and Robert Gaizauskas. 2000.Annotating Events and Temporal Information in Newswire Texts. In *LREC 2000*, pages 1287–1294.

[6] 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.

[7] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, and J.: SemEval-2007 Task 15: TempEval Temporal Relation Identification. In: Proceedings of the 4th International Workshop on Semantic Evaluations (semEval-2007), pages 75-80, Prague, ACL, June 2007.

[8] Kolya A., Ekbal A., and Bandyopadhyay S. (2010). JU_CSE_TEMP: A First Step towardsEvaluating Events, Time Expressions and Temporal Relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, ACL 2010, July 15-16, Sweden, PP. 345–350.

[9] Verhagen,M. Saur´i , R. Caselli_ , T. and Pustejovsky, J.: SemEval-2010 Task 13: TempEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 57–62

[10] Llorens, H., Saquete, E., Navarro. B.,2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles. Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 284–291, Uppsala, Sweden, 15-16 July 2010.

[11] Kolya A., Ekbal A., and Bandyopadhyay S. (2010). Event-Time Relation Identification using Machine Learning and Rules. In *Proceedings of 13th International Conference on Text, Speech and Dialogue (*TSD 2010), September 6-10, 2010, Czech Republic, PP. 114-120.

[12] Kolya A., Ekbal A., and Bandyopadhyay S. (2010). Identification of Event-Time Relation: A CRF based approach. In *Proceedings of 23rd International Conference on the Computer Processing of Oriental Languages* (*ICCPOL 2010*), July 1-3, 2010, USA, PP.63-66.

[13] Pustejovsky, J., Casta˜no, J., Ingria, R.,. Saur´ı, R,.Gaizauskas, R., Setzer, A., Katz, G. and Radev, D. *TimeML: Robust specification of event and temporal expressions in text.* In AAAI Spring Symposium on New Directions in Question-Answering, pp. 28–34, CA, 2003.

[14] Mani, I., and Wilson G. 2000. *Processing of News*. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics pp. 69–76.

[15] Boguraev, B., Ando, R-K. 2005. *TimeBank-DrivenTimeML Analysis. Annotating, Extracting and Reasoning about Time and Events* 2005.

[16] Bethard, Steven and James H. Martin. 2006. Identification of event mentions and their semantic class.In *EMNLP: Proceedings of the Conference on Empirical Methods in NLP*, pages 146–154. ACL.

[17] Mao,T.,Li.T.,Huang,D.,Yang,Y., 2006. Hybrid Models for Chinese Named Entity Recognition. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing.

[18] March, Olivia and Timothy Baldwin. 2008. Automatic event reference identification. In *ALTA 2008*,pages 79–87, Australia. International Journal of Lexicography 3(4), 235–312.

[19] B Boguraev, J Pustejovsky, R Ando, M Verhagen Timebank evolution as a community resource for TimeML parsing*Language Resources and Evaluation 41*(*1*), 91--115, Springer, 2007

[20] Pustejovsky,J. ,Hanks,P. ,Saur´I,R., See,A., Gaizauskas, R.,Setzer, A.,Radev, D.,Sundheim,B., Day,D., Ferro,L., and Lazo,M., 2003b. *The TIMEBANK corpus*. In Proceedings of Corpus Linguistics 2003, pages 647–656, Lancaster, March.

[21] Roser Saurí, Robert Knippen, Marc Verhagen and James Pustejovsky. 2005. Evita: A Robust Event Recognizer for QA Systems. Short Paper. Proceedings of HLT/EMNLP 2005: 700-707.

[22] Boguraev B. and Ando. R.K. 2005. TimeML-compliant Text Analysis for Temporal Reasoning. Proceedings of IJCAI, UK.

[23] Verhagen, M., Mani, I., Sauri, R., Littman, J., Knippen, R., Jang, S. B., Rumshisky, A., Phillips, J., & Pustejovsky, J. (2005). Automating Temporal Annotation with TARSQI'. In: 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05). Ann Arbor, Michigan, Poster/Demo.

[24] Lafferty, J., McCallum, A.K., Pereira, F., *ConditionalRandom Fields:Probabilistic Models for Segmenting and Labeling Sequence Data.* In Proceedings of 18th International Conference on Machine Learning, 2001.

[25] Sha, F., Pereira, F.,: *Shallow Parsing with Conditional Random Fields,* In: Proceedings of HLT-NAACL,2003.

[26] Vapnik, Valdimir N. 1995. *The Nature of Statistical Learning Theory*. Springer.Vendler, Zeno, 1967. *Linguistics and philosophy,chapter Verbs and times,* pages 97–121. Cornell

[27] Joachims , T. 1999. Making Large Scale SVM Learning Practical. In *B. Scholkopf, C. Burges and A. Smola editions, Advances in Kernel Methods-Support Vector Learning, MIT Press*.

[28] Miller, G.A.: WordNet: *An on-line lexical database.* International Journal of Lexicography 3(4), 235–312 (1990).

[29] Gildea, D. and D. Jurafsky. 2002. *Automatic Labeling of Semantic Roles.* Computational Linguistics, 28(3):245–288.

[30] Pradhan S., Wayne Ward, Kadri Hacioglu, James H. Martin, Daniel Jurafsky, *Shallow Semantic Parsing using Support Vector Machines* . Proceedings of HLT/NAACL-2004, Boston, MA, May 2-7, 2004.

[31] S. Pado, M. Pennacchiotti, and C. Sporleder: Towards semantic role labelling for event nominalisations through bootstrapping from verbal data. Proceedings of COLING 2008.

[32] Gurevich, O., R. Crouch, T. King, and V. de Paiva.2006. *Deverbal Nouns in Knowledge Representation.* Proceedings of FLAIRS, pages 670–675, Melbourne Beach, FL.

[33] Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik.1985. *A Comprehensive Grammar of the English Language*. Longman.

[34] Grimshaw, J. 1990. *Argument Structure. MIT Press.*

Authors

Anup Kumar Kolya is a Phd student of Computer Science and Engineering Department at Jadavpur University. He has completed his B.Tech from Kalyani University and M.E. from Jadavpur University. He is also working as Senior Research Engineer in the DIT funded project "English to Indian Language Machine Translation Phase II". His research is "Identification of Time and Sentiment features of Events and analyzing the temporal relations". He has so far published 13 papers.

Dr. Asif Ekbal is working as an Assistant Professor since 2011 in the Computer Science & Engineering Department at IIT Patna. He has completed his Ph.D degree from Jadavpur University.His research interests include Natural Language Processing, Bio-NLP, Bio-informatics. He has published a large number of journal and conference publications.

Professor Sivaji Bandyopadhyay is working as a Professor since 2001 in the Computer Science and Engineering Department at Jadavpur University, Kolkata, India. His research interests include machine translation, sentiment analysis, textual entailment, question answering systems and information retrieval among others. He is currently Supervising six national and international level projects in various areas of language technology. He has published a large number of journal and conference publications.