A Hybrid Approach for Event Extraction and Event Actor Identification

Anup Kumar Kolya[1] **Asif Ekbal**[2] **Sivaji Bandyopadhyay**[1]
[1] Computer Science and Engineering Department, Jadavpur University, India
[2] Patna (IITP), India
anup.kolya@gmail.com, asif.ekbal@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract

This paper, we propose an approach for *event extraction* and corresponding *event actor* identification within the TimeML framework. Firstly, for *event extraction*, we develop SVM based hybrid approach and for *event actor* identification the *baseline* model is developed based on the *subject* information of the dependency-parsed event sentences. Then we develop an unsupervised syntax based model that is based on the relationship of the event verbs with their argument structure extracted from the *head* information of the chunks in the parsed sentences. Evaluation on a collection of TempEval-2 corpus shows the precision, recall and F-measure values for the *baseline* model as 64.31%, 67.74% and 65.98%, respectively and the syntax based model as 69.12%, 66.90% and 67.99%, respectively.

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 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 summarization (Daniel et al., 2003) and question answering (QA). TimeML (Pustejovsky et al., 2003) presented a rich specification for annotating events in NL text extending the features of the previous one.

This paper is focused on the TimeML view of events. TimeML defines events as situations that *happen or occur*, or elements describing *states or circumstances* in which something obtains or holds the truth. These events are generally expressed by tensed or un-tensed verbs, nominalizations, adjectives, predicative clauses or prepositional phrases. The 2007 TempEval challenge attempted to address this question (Boguraev et al, 2005). In 2010, TempEval-2, event extraction task was introduced as task B. Let us consider a sentence like,

*BAGHDAD, Iraq (AP) _ an American leader of a U.N. weapons inspection team **resumed work** in Iraq Friday, nearly two months after his team was effectively **blocked.***

Sentence 1 has three events, namely '***resumed***', '***work***' and '**blocked**'. In this sentence *resumed* and *blocked* can be considered as verbal events but *work* is an nonverbal event. Generally, verbal or non-verbal event are executed by some abstract entities, directly or indirectly. Entities are basically person, organization or location.

2 Event Extraction

Below we present our hybrid approach for event extraction. The system is based on a supervised machine learner, Support Vector Machine (SVM). It makes use of the various features extracted from the TimeML corpus. In order to improve the performance of the system, we incorporate the knowledge of semantic role labeling, WordNet and several heuristics.

2.1 SVM based Approach

Initially, we started with the development of an event extraction method based on SVM. This is used as the *baseline* model. The SVM system is developed based on (Valdimir, 1995), which perform classification by constructing a N-dimensional hyperplane that optimally separates data into two categories. We use *YamCha* toolkit[1], a SVM-based tool for detecting classes in documents and formulating the event extraction

[1] http://chasenorg/~taku/software/yamcha

Proceedings of Recent Advances in Natural Language Processing, pages 592–597,
Hissar, Bulgaria, 12-14 September 2011.

task as a sequential labeling problem. Here, the *pair wise* multi-class decision method and *polynomial kernel function* are used. We use TinySVM-0.0[2] classifier for classification.

We extract the gold-standard TimeBank features for events in order to train/test the SVM model. We mainly use the various combinations of *part of speech* (PoS), *event tense*, *event aspect*, *event polarity*, *event modality*, *event stem* and *event class* features.

2.2 Use of Semantic Roles for Event Extraction

We use Semantic Role Label (SRL) (Gildea et el, 2002; Sameer et al, 2004) to identify different features of the sentences of a document. These features help us to extract the events from the text. 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 nominalisations (or, *deverbal nouns*) are commonly defined as nouns, morphologically derived from verbs, usually by suffixation (Quirk et al., 1985). Let us consider the following example sentence to understand how semantic roles can be used for event extraction.

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]

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 SVM model. But, there exists many nominalised event expressions (i.e., *deverbal nouns*) that are not identified as events by the supervised SVM. These nominalised expressions are correctly identified as events by SRL. We observe performance improvement with the inclusion of this module.

2.3 Use of WordNet for Event Extraction

WorldNet (Miller, 1990) is mainly used to identify *non-deverbal event nouns*. We observed from the outputs of SVM and SRL that the event entities like '*war*', '*attempt*', '*tour*' etc. are not properly identified. These words have noun PoS categories, and the SVM along with SRL can only identify those event words that are verbs. 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 (for example *war*) tokens having noun 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 an event.

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.

We observe significant performance improvement on event extraction with the above mentioned two rules.

2.4 Use of Rules for Event Extraction

We used WordNet to extract the 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 SVM based Stanford Named Entity (NE) tagger[3] 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 event extraction:

Rule-1: The morphologically deverbal nouns are usually identified by the suffixes like '*-tion*', '*-ion*', '*-ing*' and '*-ed*' etc. The non NE nouns but ends with these suffixes are considered as the event words.

Rule-2: After searching verb-noun combination from the test set, non-NE noun words are considered as the events.

Rule- 3: The non-NE nouns occurring after (i) the complements of aspectual PPs headed by prepositions (ii) any time-related verbs (iii) certain expressions are considered as events.

[2](http://cl.aist-nara.ac.jp/~taku ku/software/TinySVM)

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

2.5 Evaluation Results

We use the TempEval-2010 datasets to report the evaluation results. We develop a number of SVM models depending upon the various features included into it. 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. We used different feature combinations within the context of previous 3 and next 3 words. The test data had 373 verbal and 125 nondeverbal event nouns. Overall evaluation results are reported in Table 1. The SVM based system shows the precision, recall and F-measure values of 75.8%, 78.5% and 77.13%, respectively. The performance increases by almost 1.59 percentage F-measure points with the use of semantic roles. Table 1 shows very high performance improvement (i.e., 10.98%) with the use of WordNet. The rule-based component also shows the effectiveness with the improvement of 5.37 F-measure percentage points. Finally, the system achieves the precision, recall and F-measure values of 93.00%, 96.00% and 94.47%, respectively. This is actually an improvement of approximately 12% F-measure value over the best reported system.

3 Event Actor Identification

In this section, we detail our method for event actor identification.

3.1 Subject Based Baseline Model

We have preprocessed the TempEval-2 corpus for identifying the actors of the events. We have previously (Kolya et al. 2010a; Kolya et al 2010b) worked on the various problems of event and temporal relation identification such as (i). Event-time and event-documentation creation time (DCT) temporal relation (TE) identification in the same sentence, (ii). Even–event temporal relation identification in two consecutive sentences and (iii). Subevent-subevent temporal relation identification in the same sentence on the Tempeval-1 and Temeval-2 corpus. We have observed from this experience that almost all events are involved with the actors, either active or passive. Actually, event actions are done by someone or somebody is doing this kind of action. Event actions involve with person, organization and sometimes with location also. In the present attempt, we consider the approaches that were conducted for identifying emotion holders (Das and Bandyopadhyay, 2010). Thereafter, we came up with the following heuristics for actor identification, (i) we discard the non-event sentences, i.e. those sentences that don't contain any event entity. (ii) If multiple events exist in any sentence, then all the events will have the same actors. Once an actor is identified for any event, it is assigned to the other event as well. (iii) If there are multiple actors and events, then <event, actor> pairs are formed by considering an event and its closest possible actor in the sentence. All the events may not have an active actor. The actor may be passive also. For example, consider the following sentences:

Table 1. Evaluation results of event extraction

Model	precision	recall	F-measure
SVM	75.80	78.50	77.13
SVM+SRL	77.20	80.30	78.72
SVM+SRL+WordNet	89.30	90.10	89.70
SVM + SRL + WordNet + Rules	93.50	96.70	95.07

1. This time a <bomb/> at an abortion clinic.

2. *Plates <recovered/> at the Olympic park bombing <appear/> to <match/> those <found/> at the abortion clinic <bombing/> in Atlanta.*

Corpus Preparation: We did not have any gold standard corpus for event actor identification. We have used the Temeval-2 corpus as a gold standard event actor corpus by manually annotating event actors in each sentence. The gold corpus looks as follows:

<eActor> People </eActor> have <predicted/> his <demise/> so many times , and the <eActor>US<eActor> has <tried/> to <hasten/> it on several occasions .

Here, a "*People*" is the event actor of both events **predicted** and **demise,** and "*US*" is the event actor of the events, **tried** and **hasten.** This corpus has 11 documents, 156 sentences and 459 events.

3.2. Baseline Model based on Dependency Parsing and Subject Extraction

Stanford Parser (de Marneffe et al,2006), a probabilistic lexicalized parser containing 45 differ-

ent PoS tags of Pen Tree bank is used to get the parsed sentences with dependency relations. The input event sentences are passed through the parser. The dependency relationships extracted from the parsed data are checked for predicates "*nsubj*" and "*xsubj*" so that the *subject* related information in the "*nsubj*" and "*xsubj*" predicate are considered as the probable candidate for identifying the event actor. Other dependency relations are filtered out from the parsed output. The present system is developed based on the filtered subject information only. An example sentence is noted below whose parsed output and dependency relations are shown. Here, the "*nsubj*" relations containing the event word "endures" tags "eActor" as an event actor.
"*Time and again, he* ***endures.***"

```
(ROOT (S  (S  (UCP  (NP (NNP
Time))(CC and)(ADVP (RB again))))(,
,) (NP (PRP he)) (VP (VBZ endures))
(. .)))
```

```
ccomp  (endures-6,  Time-1),advmod
(Time-1, again-3),conj and (Time-1,
again-3),ccomp (endures-6, again-3)
nsubj (endures-6, he-5)
```

This *baseline* model is evaluated on the gold standard holder annotated an emotional sentence that has been extracted from the VerbNet. Total 156 sentences are evaluated and evaluation results are presented in Table 2. So, the next step is to explore the syntactical way for identifying argument structure of the sentences for their corresponding emotional verbs and to capture the emotion holder as a *thematic role* respectively.

3.3. Syntax Based Model

The syntax of a sentence is an important clue to capture the event actor inscribed in text. More specifically, the argument structure or subcategorization information for a verb plays an important role to identify the event actor from an event sentence. A subcategorization frame is a statement of what types of syntactic arguments a verb (or an adjective) takes, such as objects, infinitives, that-clauses, participial clauses, and subcategorized prepositional phrases (Manning et al. 1993). VerbNet (Kipper-Schuler et al, 2005) is the largest online verb lexicon with explicitly stated syntactic and semantic information based on Levin's verb classification (Levin et al 1993). It is a hierarchical domain-independent, broad-coverage verb lexicon with mappings to other lexical resources such as WordNet (Miller et al, 1990), XTAG (2001) and FrameNet (Baker et al, 1998). We use VerbNet throughout this experiment for identifying the event actors. The existing syntax for each event verb is extracted from VerbNet and a separate rule based argument structure acquisition system is developed in the present task for identifying the event actor. The acquired argument structures are compared against the extracted VerbNet frame syntaxes. If the acquired argument structure matches with any of the extracted frame syntaxes, the event actor corresponding to each event verb is tagged with the actor information in the appropriate slot in the sentence.

Syntax Acquisition from VerbNet: VerbNet associates the semantics of a verb with its syntactic frames and combines traditional lexical semantic information such as thematic roles and semantic predicates, with syntactic frames and selectional restrictions. Verb entries in the same VerbNet class share common syntactic frames, and thus they are believed to have the same syntactic behavior. The VerbNet files containing the verbs with their possible subcategorization frames and membership information are stored in XML file format.

<THEMROLES/> <FRAMES>
<FRAME> <DESCRIPTION descriptionNumber="8.1" primary="TO-INF-SC"
secondary="" xtag="0.1"/> …. <EXAMPLE>I loved to write.</EXAMPLE>
<SYNTAX> <NP value="Experiencer">
<SYNRESTRS/> </NP>
<VERB/> <NP value="Theme">
<SEMANTICS> <PRED value="**event_state**">
<ARGS> <ARG type="Event" value="E"/>
<ARG type="**VerbSpecific**"
value="**Event**"/> <ARG type="ThemRole" value="Passive"/> …..
</ARGS> </PRED> </SEMANTICS>
</FRAME>….

The XML files of VerbNet are preprocessed to build up a general list that contains all member verbs and their available syntax information retrieved from VerbNet. This preprocessed list is searched to acquire the syntactical frames for each event verb. One of the main criteria considered for selecting the frames is the presence of "*event_state*" type predicate associated with the frame semantics.

Argument Structure Acquisition Framework: To acquire the argument structure for a

sentence, two separate approaches, Methods A and B, have been used, one (Method A) is from the parsed result directly and another (Method B) is from the PoS tagged and chunked sentences accordingly. The parsed event sentences are passed through a rule based *phrasal-head* extraction process to identify the phrase level argument structure of the sentences corresponding to the event verbs. The extracted *head part* of every phrase from the well-structured bracketed parsed data is considered as the component of the argument structure. For example, the *head* parts of the phrases are extracted to make the phrase level pattern or argument structures of the following sentences.

Sentence1: "*Ram killed Shyam with a knife.*"
Parsed Output:
(ROOT (S (**NP** (NNP Ram)) (**VP** (VBD killed) (**NP** (NNS Shyam)) (**PP** (IN **with**) (NP (DT a) (NN knife)))) (. .)))
Acquired Argument Structure: [NP VP NP PP-with]
Simplified Extracted VerbNet Frame Syntax: [<NP value="*Actor*"> <VERB/> <NP *patient*> <PREP value="*with*">]

Event Actor for Event Verbs–The role of Subject and Syntax: It is to be mentioned that the phrases headed by "S" (sentential complement), "PP" (Preposition Phrase), "NP" (Noun Phrase) followed by the event verb phrase contribute in structuring the syntactical argument. One tag conversion routine has been developed to transform the POS information of the system-generated argument structure for comparison with the POS categories of the VerbNet syntax. It has been observed that the phrases that start with ADJP, ADVP (adjective, adverbial phrases) tags generally do not contribute towards valid argument selection strategy. But, the entities in the slots of active frame elements are added if they construct a frame that matches with any of the extracted frames from VerbNet. The *head* part of each phrase with its component attributes (e.g. "*with*" component attribute for "PP" phrase) in the parsed result helps in identifying the maximum matching possibilities. Another alternative way to identify the argument structure from a sentence is carried out based on the PoS tagged and chunked data. The PoS tagged sentences are passed through a Conditional Random Field (CRF) based chunker (Phan et al, 2006) to acquire chunked data where each component of the chunk is marked with *beginning* or *intermediate* or *end* corresponding to the elements slot in that chunk. The POS of the *beginning* part of every chunk are extracted and frames are developed to construct the argument structure of the sentence corresponding to the event verb. The acquired argument structure of a sentence is mapped to all of the extracted VerbNet frames. If a single match is found, the slot devoted for the actor in VerbNet frame is used to tag in the appropriate slot in the acquired frame. For example, the argument structure acquired from the following chunked sentence is "NP-VP-NP".

But, it has been observed that this second system suffers from the inability to recognize arguments from adjuncts as the system blindly captures *beginning* parts as arguments whereas they are adjuncts in real. So, this system is biased to the *b ginning* chunk.

3.4. Evaluation

The evaluation of the baseline system is straightforward. The event actor annotated sentences are extracted from the VerbNet and the sentences are passed through the baseline system to annotate the sentences with their *subject* based actor tag accordingly. Evaluation with 156 sentences is shown in Table 2. It is observed that the *subject* information helps in identifying event actor with high *recall*. But, the actor identification task for passive sentences fails in this *baseline* method and hence there is a fall in *precision* value. Two types of unsupervised rule based methods have been adopted to acquire the argument structure from the event sentences. It has been observed that, the Method-A that acquires argument structure from parsed result directly outperforms the Method-B that acquires these structures from PoS tagged and chunked data. The *recall* value has decreased in Method-B as it fails to distinguish the arguments from the adjuncts. The event actor identification system based on argument structure directly from parsed output gives satisfactory performance.

Table 2. Evaluation results of actor identification

Type	Baseline Model	Syntactic Model	
		Method A	Method B
Precision	64.31	69.12	64.05
Recall	67.74	66.90	65.52
F-measure	65.98	67.99	64.78

4 Conclusion

In this paper, we have reported our work on event extraction under the TempEval -2010 evaluation exercise. Initially, we developed a SVM based supervised system in conjunction with number of techniques based on SRL, WordNet and handcrafted rules for event extraction. We then identify the actors for the events based on the roles associated to *subject* information. The syntactic way of developing the actor extraction module by focusing on the role of arguments of the event verbs improves the result significantly.

Future works include the identification of more precise rules for event identification and multiword events. The actor-annotated corpus preparation from VerbNet especially for event verbs followed by the argument extraction module can be further explored through the help of machine learning approach.

Acknowledgments

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

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

Baker, C.F., Fillmore, C.J., Lowe, J.B.: *The Berkeley FrameNet project.* COLING/ACL, pp. 86–90 (1998)

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

Das Dipankar and Sivaji Bandyopadhyay. 2010. Emotion Holder for Emotional Verbs–The role of Subject and Syntax. CICLing- 2010), A. Gelbukh (Ed.), LNCS 6008, pp. 385-393, Romania

De Marneffe, M.-C., MacCartney, B., Manning, C.D.: *Generating Typed Dependency Parses from Phrase Structure Parses.* LREC (2006)

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

Kipper-Schuler, K.: VerbNet: *A broad-coverage, comprehensive verb lexicon.* Ph.D. thesis,Computer and Information Science Dept., University of Pennsylvania, Philadelphia,PA (2005)

Kolya, A., Ekbal, A. and Bandyopadhyay, S. 2010. JU_CSE_TEMP: *A First Step towards Evaluating Events, Time Expressions and Temporal Relations.* SemEval, ACL, July 15-16, Sweden, pp. 345–350.

Kolya, A., Ekbal, A. and Bandyopadhyay, S. (2010a). *Event-Time Relation Identification using Machine Learning and Rules.* In Proceedings of 13th International Conference on Text, Speech and Dialogue, 2010, pp. 114-120.

Kolya, A., Ekbal, A. and Bandyopadhyay, S. (2010b). *Identification of Event-Time Relation:* A CRF based approach. ICCPOL 2010, USA, PP.63-66.

Levin, B.: English Verb Classes and Alternation: *A Preliminary Investigation.* The University of Chicago Press, Chicago (1993)

Manning, C.D.: *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora.*In: 31st Meeting of the ACL, Columbus, Ohio, pp. 235–242 (1993)

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

Pustejovsky, James, Jos´e M. Casta˜no, Robert Ingria,Roser Saur´?, Robert Gaizauskas, Andrea Setzer, andGraham Katz. 2003. *TimeML: Robust Specification of Event and Temporal Expressions in Text.* In IWCS-5.

Phan, X.-H.: CRFChunker: *CRF English Phrase Chunker.* In: PACLIC 2006 (2006)

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

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