NATURAL LANGUAGE SEMANTIC EVENT EXTRACTION PIPELINE

Siaw Nyuk Hiong[1], Narayanan Kulathuramaiyer[2], and Jane Labadin[3]
University Malaysia Sarawak, Malaysia,
[1]ftsm2006@yahoo.com , [2]nara@fit.unimas.my, [3]ljane@fit.unimas.my

ABSTRACT. This paper presents a novel contribution of this research which is an automated NLP pipeline for semantic event extraction and annotation (EveSem). The output from this research is an xml annotated semantic events. Temporal interpretation of event is incorporated by using the linguistic elements made available through the use of the tools. A preliminary evaluation showed that EveSem performed equally well as TIPSem in extracting verbal event with a precision of 85.42 and a recall of 89.13. This work can contribute towards automated annotation of semantic event corpus and event timeline construction as future research.

Keywords: natural language processing, information extraction, event extraction, semantic event

INTRODUCTION

Information extraction (IE) is becoming an important task for knowledge acquisition from natural language text (Berrazega, 2012). Different definitions of IE are (a) IE as a task to obtain structured information from unstructured text (Piskorski &Yangarber, 2013) and (b) IE as "automatic extraction of structured information such as entities, relationships between entities and attributes describing entities from unstructured sources" (Chambers, 2011). IE can be a component in natural language processing (NLP) applications such as Machine Translation, Question Answering, Text Summarization, Opinion Mining, etc (Piskorski & Yangarber, 2013).

Event extraction is high level IE task. Automated event extraction identifies and labels semantic roles of event such as participant, location and time (Liu et al., 2013). Furthermore, semantic representation of events help to interprete "Who did what to whom when where?" of text (Filatova & Hatzivassiloglou, 2003; Llorens, Navarro & Saquete, 2009; Piskorski & Yangarber, 2013). Message Understanding Conferences ([1]MUC, 1987-1997) and [2]ACE program (since 2000) are two important platforms to study IE. However, MUC event extraction is domain dependent and labour intensive in creating the event templates. Furthermore, these fixed templates are not able to detect multiple-type event. Other information extraction researches to extract semantic events have been carried out by Surdeanu et al. (2003), McCracken (2006), Ji and Grishman (2008) and Wang (2012). Semantic event extraction is also an area of active research in other fields like text summarization (Filatova & Hatzivassiloglou, 2004; Li et al., 2006; Liu et al., 2007) and text mining (Capet et al., 2008; Cohen et al., 2009; Hung et al., 2010; Llorens et al., 2010). Data experimentation from these researches have shown positive results. Many natural language processing (NLP) tools for part of speech (POS) tagging, name entity recognition (NER), lexical databases like PropBank (Palmer, Gildea & Kingsbury, 2005), VerbNet (Kipper-

[1] http://www-nlpir.nist.gov/related_projects/muc/.
[2] http://www.itl.nist.gov/iad/mig/tests/ace/.

Schuler, 2005) and lexical semantic like WordNet (Miller, 1995) have been used to process the text for semantic event extraction in these researches. Furthermore, conventional techniques that extract named entities and events in text are not able to detect temporal relationship between events and their chronological ordering (Berrazega, 2012).

In addition to that, there is no semantically annotated event corpus for natural language text which can be used for the purpose of research. TimeBank (Pustejovsky, 2006) is a corpus with annotated event but it is not semantically annotated with event participants (UzZaman et al., 2013). There is a need first to research on the feasibility of automatically annotating semantic events in a raw text with the many available NLP tools. Many robust, efficient and high-coverage shallow text processing techniques have been applied in IE to process textual data (Piskorski & Yangarber, 2013). Since semantic event extraction using natural language approaches have been widely researched on (Surdeanu et al., 2003; Filatova & Hatzivassiloglou, 2004; Li et al., 2006; McCracken, 2006; Liu et al., 2007; Capet et al., 2008; Ji & Grishman, 2008; Cohen et al., 2009; Hung et al., 2010; Llorens et al, 2010; Wang, 2012), it is feasible to study NLP approaches to automate semantic event extraction and annotation. The needs of automatically annotating raw text with semantic events have generated a research gap which formulated the motivation for this research. The main aims of this research are to (a) propose and (b) implement NLP approaches to automate semantic event extraction and annotation.

RELATED RESEARCH

Semantic event extraction had been carried out in the fields of information extraction (MUC; ACE; Surdeanu et al., 2003; McCracken, 2006; Ji & Grishman, 2008; Wang, 2012), text summarization (Filatova & Hatzivassiloglou, 2004; Li et al., 2006; Liu et al., 2007) and text mining (Capet et al., 2008; Cohen et al., 2009; Hung et al., 2010; Llorens et al, 2010). Both MUC and ACE program used pattern-based approach with predefined templates for events. Surdeanu et al. (2003) applied semantic role label to parse a text into predicate and argument structure used to match and fill the slots specified for the event. McCracken (2006) used PropBank semantic role label to represent factual event in the text and used in frame representation of the event whereas Ji and Grishman (2008) use statistical information of triggers and arguments associated with events to corrent event and argument identification and classification for ACE evaluation. Wang (2012) used both semantic role label (SRL) and NER method to extract semantic events from Chinese news based on ACE definition.

The event oriented approach is mostly applied in the field of text summarization. Semantic features in the form of event terms (Liu et al., 2007) and named entities with relation (Filatova and Hatzivassiloglou, 2004) were used for clustering. Liu et al. (2007) constructed event terms graph with reference to VerbOcean. Experimentation result indicates that event-based features gave better summarization result compared to word features (Filatova and Hatzivassiloglou, 2004). Li et al. (2006) used terms with associated entities as the semantic event terms.

In the field of text mining, a hybrid approach for semantic event extraction was used (Capet et al., 2008; Cohen et al., 2009; Hung et al., 2010). Capet et al. (2008) used lexico-semantic patterns for concept matching using dependency chains. Cohen et al., 2009 extracted medical events by using concept recognizer on a biological domain whereas Hung et al. (2010) used semantic roles to label semantic events for sentences collected from the web using lexico-syntactic patterns. Llorens et al (2010) applied Conditional Random Fields machine learning method for event recognition and classification. Morpho-syntactic features, WordNet-based features and semantic role features were used for learning. Data experimentations indicated that semantic features could improve the event recognition and classification tasks.

The various NLP processing tasks that had been used by these researchers were:

(a) POS and syntactic parsing: Capet et al. (2008), Hung et al. (2010), Llorens et al. (2010)
(b) NER: MUC, ACE, Filatova and Hatzivassiloglou (2004), Li et al. (2006), Liu et al. (2007),
(c) SRL: Surdeanu et al. (2003), McCracken (2006), Hung et al. (2010), Llorens et al. (2010), Wang (2012).

Beside using SRL, some researchers had applied other lexical resources like VerbOcean (Liu et al., 2007), name entities relationship (Filatova & Hatzivassiloglou, 2004; Li et al., 2006), domain concept (Cohen et al., 2009) and dependency chains (Capet et al., 2008) to represent and extract semantic events. Most of the text used in these researches were of generic domain (Surdeanu et al., 2003; Filatova & Hatzivassiloglou, 2004; Li et al., 2006; Liu et al., 2007; Hung et al., 2010; Llorens et al, 2010; Wang, 2012).It can be concluded that NLP tools played a very important role in extracting semantic events for all these researches. This includes the use of NLP resources like PropBank, VerbNet, WordNet and VerbOcean.

TEMPORAL INTERPRETATION OF EVENT

From linguistic perspective, three main components involved in temporal interpretation of event are situation type, tense and aspect. Situation type describes the semantic content of verb as state, activities, achievements or accomplishment (Vendler, 1967). Both tense and aspect relate closely to time. The interaction of the two elements mapped to the past, present and future with reference to speech time (Reichenbach, 1947) on a timeline is shown in Figure 1. In the present perfect, event time (ET) precedes the speech (ST) and reference time (RT) in which reference time is the same as event time (ET<ST=RT). In the past perfect, the relationship between ET, ST and RT is ET<RT<ST. For future perfect, ST<ET<RT. In the progressive, it indicates the time of core event happening without knowing the resultant state. Table 1 shows a summary of the three components that have been described.

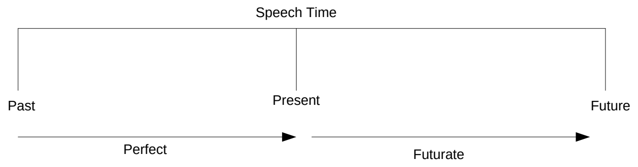


Figure 1. Relationship Between Tense And Aspect On A Timeline

Table 1. Three Components Of Temporal Interpretation For Event

Event Temporal Interpretation Components		
Situation type (Vendler, 1967)	Tense	Aspect
State	Past	Perfective
Activities	Present	Progressive
Achievements	Future	
Accomplishments		

PROPOSED WORK

This research proposed the application of NLP tools as a pipeline for semantic event extraction and annotation (EveSem) of a generic raw text. The Semantic parsing: PropBank-NomBank frames developed by [3]Lund University is selected as the SRL package. This tool can extract semantic events for both verb and noun which make it an advantage over other

[3] http://nlp.cs.lth.se/software/semantic_parsing%3A_propbank_nombank_frames/.

tools. The semantic parsing of the tool is trained on both PropBank and NomBank. The predicate-arguments of the verbs and nouns are extracted through a [4]Semantic-Parsing interface. Semantic role of extracted verbal predicate for PropBank is mapped to VerbNet semantic role through the use of [5]SemLink (Rodney, 2008). Frame semantic of the mapped verb is extracted to obtain the semantic representation in [6]VerbNet3.2. The frame semantic of VerbNet contains situation type which is one of the components for temporal interpretation of event. The tense and aspect components can be obtained from the part of speech for each verb. Java is the programming tool used to implement the pipeline. Figure 2 shows the conceptual framework for the proposed work and Table 2 shows the input-process-output involved in each implementation stage of the pipeline.

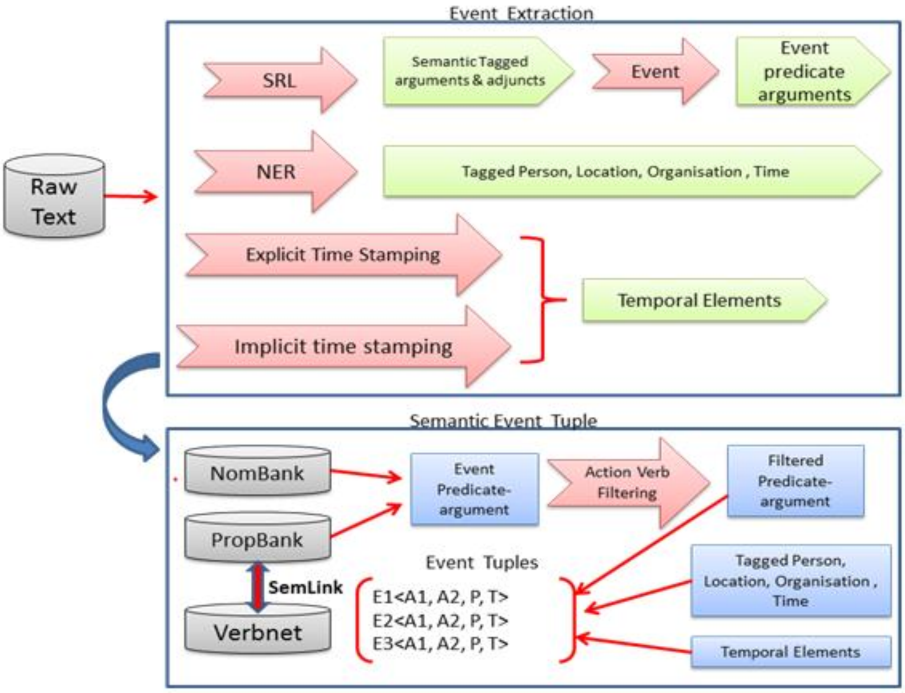


Figure 2. Natural Language Semantic Event Extraction Annotation (EveSem) Pipeline

Table 2. EveSem Implementation

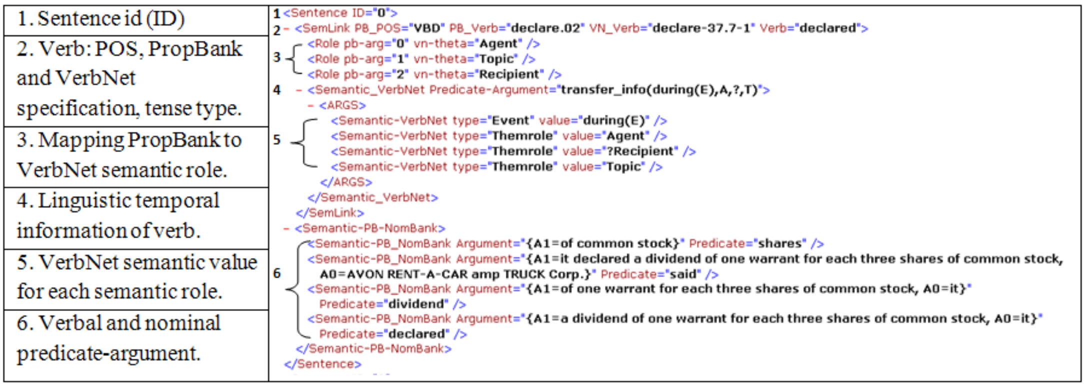
NLP Tools/Resources	Input	Process	Output
[3]Semantic parsing: PropBank--NomBank frames	raw text	(1) *scripts/preprocess.sh* - tokenizes, adds part-of-speech tags, and finds lemmas. (2) *scripts/run.sh* - second-order dependency parser, linguistic constraints, semantic reranking, and syntactic–semantic integration	CoNLL-2008 format, e.g. [0, Pacific, _, _, NNP, Pacific, _, NNP, 3, NAME, _, _, _, _] [1, First, _, _, NNP, First, _, NNP, 3, NAME, _, _, _, _]
[4]Semantic-Parsing interface	CoNLL-2008 format	Extract predicate-argument for verb and noun event trigger.	Predicate-argument, e.g. protect{A1=shareholders, A0=it} priced{A1=an inadequately priced takeover offer}

[3] http://nlp.cs.lth.se/software/semantic_parsing%3A_propbank_nombank_frames/
[4] https://github.com/xiejuncs/Semantic-Parsing-Interface
[5] http://verbs.colorado.edu/semlink/
[6] http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html
[7] http://www.oracle.com/technetwork/java/javase/downloads/jdk6downloads-1902814.html

[5]SemLink 1.2.1	PropBank verb specification	Map PropBank verb to VerbNet	VerbNet Semantic role label.
[6]VerbNet 3.2	Verb specification	Extract linguistic temporal information and value for each semantic role of verbal event.	Predicate-argument of linguistik temporal information and value for each semantic role.
Programming Tool			
[7]Java (JDK 6)	Output from NLP tools and resources	Process output from the NLP tools and NLP resources in a pipeline manner.	XML annotated semantic events (Figure 3).

SEMANTIC EVENT ANNOTATION

The extracted events are annotated accordingly as : (1) sentence ID, (2) POS for verb, PropBank and VerbNet verb specification, tense type, (3) PropBank and VerbNet semantic role label, (4) predicate-argument of linguistic temporal information, (5) VerbNet semantic value for each semantic role, (6) predicate-argument of each verb and noun. The annotated text is output as an xml file which can be retrieved for further processing later. Figure 3 shows an example of the xml semantic event annotation.



1. Sentence id (ID)
2. Verb: POS, PropBank and VerbNet specification, tense type.
3. Mapping PropBank to VerbNet semantic role.
4. Linguistic temporal information of verb.
5. VerbNet semantic value for each semantic role.
6. Verbal and nominal predicate-argument.

```
1 <Sentence ID="0">
2 - <SemLink PB_POS="VBD" PB_Verb="declare.02" VN_Verb="declare-37.7-1" Verb="declared">
      <Role pb-arg="0" vn-theta="Agent" />
3     <Role pb-arg="1" vn-theta="Topic" />
      <Role pb-arg="2" vn-theta="Recipient" />
4   - <Semantic_VerbNet Predicate-Argument="transfer_info(during(E),A,?,T)">
     - <ARGS>
          <Semantic-VerbNet type="Event" value="during(E)" />
5         <Semantic-VerbNet type="Themrole" value="Agent" />
          <Semantic-VerbNet type="Themrole" value="?Recipient" />
          <Semantic-VerbNet type="Themrole" value="Topic" />
       </ARGS>
     </Semantic_VerbNet>
   </SemLink>
 - <Semantic-PB-NomBank>
    <Semantic-PB_NomBank Argument="{A1=of common stock}" Predicate="shares" />
    <Semantic-PB_NomBank Argument="{A1=it declared a dividend of one warrant for each three shares of common stock,
6     A0=AVON RENT-A-CAR amp TRUCK Corp.}" Predicate="said" />
    <Semantic-PB_NomBank Argument="{A1=of one warrant for each three shares of common stock, A0=it}"
      Predicate="dividend" />
    <Semantic-PB_NomBank Argument="{A1=a dividend of one warrant for each three shares of common stock, A0=it}"
      Predicate="declared" />
   </Semantic-PB-NomBank>
 </Sentence>
```

Figure 3. Semantic Event Annotation

TEMPORAL PROCESSING

The three components for temporal interpretation described earlier can be obtained from the semantic frame of VerbNet and the part of speech for each verb. The semantic frame of VerbNet has classification for event as: (1) states: (E), (2) activities: during(E), (3) accomplishments: result(E), (4) achievements: during(E), end(E). Temporal information infers from these components can be used to augment other explicit temporal information for event timeline construction later on. Figure 4 shows the components for event temporal interpretation.

Event Temporal Interpretation Components		
Situation type (Vendler, 1967)	Tense	Aspect
verbNet Semantic Frame <Semantic_VerbNet Predicate-Argument="transfer_info(**during(E)**,A,?,T)">	Annotation: Part of speech (POS) <SemLink PB_**POS="VBD"** PB_Verb="say.01" VN_Verb="say-37.7-1" Verb=" said">	

Figure 4. Temporal Interpretation Components

PRELIMINARY EVALUATION

A total of 12 World Street Journal (WSJ) news articles from [8]TimeBank1.2 were selected and used as the raw texts for this evaluaton. TIPSemB-1.0 (Llorens et al., 2010) was used as the reference event annotation system. It topped the performance for Task B (event extraction and classification) of TempEval-3 evaluation (UzZaman et al., 2013). Precision and recall metrics were used to measure the performance. Similar metrics had been used in TempEval-3 evaluation and the equations are:

$$\text{Precision} = \frac{[Sysentity \cap \text{Refentity}]}{Sysentity} \tag{1}$$

$$\text{Recall} = \frac{[Sysentity \cap \text{Refentity}]}{\text{Refentity}} \tag{2}$$

where *Sysentity* are entities extracted by the system to evaluate and *Refentity* are entities from the reference annotation that are being compared.

A total of 151 verbal and 127 nominal events were extracted by EveSem. The overall results showed that EveSem could perform well in extracting verbal events with a precision of 85.42 and a recall of 89.13. On the contrary, extracting nominal events gave a precision of only 24.90 and a recall of 63.40. TIPSem had used PropBank SRL features besides WordNet lexical semantics features and other morphosyntactic features for machine learning. This could explain for the high precision and recall obtained for verbal event extraction in EveSem which also used PropBank. EveSem and TIPSem used different approach in extracting nominal events. EveSem used NomBank features whereas TIPSem used a combination of lexical, semantic role and other morphosyntactic features. Even though EveSem extracted nominal event differently, it can still complement TIPSem in extracting semantic events. EveSem's contribution towards automatic semantic event extraction and annotation (verbal and nominal events) proved to be novel and significant.

CONCLUSION

The novel contribution from this research is a NLP pipeline for the automation of: (a) semantic event extraction, and (b) semantic event annotation (EveSem). It helps to close the gap that formulated the research motivation and the aims of the research have been achieved. Future research can lead to automated annotation of semantic event corpus and timeline construction. It has potential for research in other field as well, for example smart communities. In this context, news reader that extract and generate important and related news requested by a user using semantic event features can be carried out. In e-learning, a semantic event corpus can aid in teaching text summarization and paraphrasing. As a conclusion, this research has paved the way towards automated semantic event extraction and annotation which opens up more research opportunities in various fields.

REFERENCES

Chambers, N. (2011). *Inducing event schemas and their participants from unlabeled text*. (Doctoral Dissertation, Stanford University).

Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon.* (Doctoral Dissertation, University of Pennsylvania).

[8] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T08

Llorens, H., Saquete, E., & Navarro, B. (2010). TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation,* 284-291. Uppsala, Sweden: Association for Computational Linguistics.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM. 38(11)* , 39-41.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotation corpus of semantic roles. *Computational Linguistics 31(1).*

Piskorski, J. & Yangarber, R. (2013). Information extraction: past, present and future. In T. P. (Eds), *Multi-source, multilingual information extraction and summarization 11, theory and application of natural langauge processing.* Berlin: Springer-Verlag.

Reichenbach, H. (1947). *Elements of Symbolic Logic.* London. UK: Macmillan.

Rodney, D., Wayne, W., James, H., & Palmer, M. (2008). Extracting a representation from text for semantic analysis. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, (pp. 241-244).

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., & Pustejovsky, J. (2013). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events and temporal relations. *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013).*

Vendler, J. (1967). *Linguistics in philosophy.* Ithaca, N.Y.: Cornell University Press.