International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.163.01

# Event Extraction and Representation Model from News Articles

Bekele Abera Hordofa
*College of Science and Technology*
*Oromia State University, Batu, Oromia, Ethiopia*

**Abstract: Events are dynamic data structures that play a key role in understanding phenomena happening in real world, which are basically driven by the 4Ws' (what, who, when, and where). This natural evolution of questions is a typical example of what one might ask about an event. Event is a natural way to explain complicated relations between people, places, actions and objects. Event centered modeling captures the dynamic aspects of an event along with semantic representation of event facts. In this research work, the researcher proposed event extraction and representation model from news articles. Event modeling involves seminal event identification, event elements extraction and event semantic elements representation. Maximum entropy classifier is trained and tested using event related Amharic news articles collected from archive of Fana Broadcast Corporate. For event representation, the researcher designed ontology based event representation model that provides deeper semantic through event information representation. Evaluation of trigger identification and event elements extraction is carried out by comparing manually tagged news articles with the automatically extracted event information by the system. The evaluation result shows that the trigger identifier module obtains precision (67.1%) of event correctly which contributes to the better event elements extraction. The event elements extractor component shows greater obtaining precision (69.1%) while event classification module classify about (72%) of event correctly. The representative ability of our event representation model is evaluated with respect to requirements and event dimensions covered in this work.**

**Keywords: Event, Event Modeling, Event Trigger, Event Elements, Event Representation**

## I. INTRODUCTION

The notion of event has been widely used in many research fields like natural language processing, information retrieval and information extraction [1]. There is no generally accepted definition of an event. Event is something unusual, fresh, something that people have not heard before and crucially, is of interest to readers like car accident, conference, natural disaster, *etc.* According to Miller *et al.* [2], event is something that happens or occur at a given place and time. In topic detection and trucking, an event is defined as something that happens at a given place and time, along with all the necessary preconditions and unavoidable consequences [3].

In todays real world, a number of planned and unplanned events can happen or occur every day. These events are presented or published in the form of news over news broadcast services or on social media networks, that is unstructured and heterogeneous events over dynamic web. At the same time, the number of internet users, web news articles that are writen in various language online news broadcast, micblog service providers, *etc.*, are leading to information explosion over the web. A large percentage of these articles are discussing current, past and future real world events. Automatically identifying news article associated with seminal events is a challenging problem due to the heterogeneous and unstructured nature of news articles. Seminal event is something new wich is interesting to public. Seminal events are more accessible and understandable, if associated event information are extracted and semantically represented.

Event extraction is high level information extraction task which tries to formulate an event as who did what to whom, when and where [1, 4]. Most news and microblogs in the world give priority to events in their country, unless a very important event happens in another part of the world, because people want to learn what is happening around them [3]. Event extraction and representation from a variety of media (news, social media) is highly demanding in government, news publishers, and other service provider organizations. One of the recognized causes for this issue is the semantic gap existing between our conceptualizations of the world, usually expressed using language or other high level abstractions, and our experience of the world [1]. People observe and understand the world through event because it is a suitable unit in accordance with aspect of human perception. Several practical applications can arise from a good solution to event modeling. Currently, this task is performed manually by media analysts, and online digital news editor, who have the task of collecting, interpreting, and presenting news from multiple news sources. In addition, a system that could recpgnize events automatically would be usefull for news applications where decision making process is based on new events and the evolution of existing events.

Event scenario is described using event elements. Event elements provide multi-dimensional semantic understanding of event description. Event extraction is langauge depedent, that is event extraction system developed for English or any other language cannot work for Amharic language of the same domain. Taking language issue into consideratio, a number of event extarction model has been developed in a variety of language such as English[5], Chinise[1, 6], Turkish[3] and others.

## II. RELATED WORK

A number of research works have been conducted on event extraction from various sources for different target applications. Design and usage of source (Wikipedia, news article, social media or other) has its own impact on the event extraction process. The task of event extraction differs according to the type of event, detection task and detection methods [14]. Event modeling is language dependent; one should take language features and encoding techniques into consideration. Development of language specific event modeling requires efficient NLP tools (Part of speech tagger, morphological analyzer, named entity recognizer and others). Amharic is one of morphologically rich language, has its own character, numbering, way of writing and encoding. Moreover, language specific issues like normalization (character, and temporal), use of NLP tools should be considered.

Fabian, et al. [7] proposed YAGO, a large event ontology constructed from Wikipedia and WordNet. Event facts are automatically extracted from Wikipedia and unified with WordNet, using combination of rule based and heuristic methods. The authors used the English version of Wikipedia in January 2007, which comprised 1.6 million articles and they considered each Wikipedia article as a single Web page and usually describes a single topic. Another works done by James, et al. [8] used Wikipedia as a target source of event. The focus of James, et al. [8], is to extract historical events from Wikipedia articles that are available for about 2,500 years for different languages. For event identification they used DBpedia ontology type event. Their event extraction is based on semantic parsing from Wikipedia text and converted them into the LODE event model.

Axel et al. [9] and Hila et al. [10], tried to extract event from social media. Hila et al. [10] present a query-oriented solution for retrieving social media documents for planned events across different social media sites. The authors used user contributed web data containing information about event. This information ranges from known event features (title, time, location) posted on event aggregation platforms to discussions and reactions related to events shared on different social media sites. Axel et al. [9] proposed event recommender system for Twitter users. It identifies twitter activity co-located with previous events, and uses it to drive geographic recommendations via item-based collaborative filtering. They used Eventbrite popular event organization website which contains a database of previous and upcoming events as source of event.

Jakub et al [11) proposed a NEXUS (News cluster Event eXtraction Using language Structures), which uses Ontology of Politically Motivated Violent Events (PMVE). NEXUS selects security related events via application of keyword based heuristics. The extracted information about violent events and related entities such as people and organizations is mapped to domain ontology PMVE. Wang et al. [1] proposed a framework that extracts event semantic elements (5W1H) from Chinese news for event ontology population. The researcher designed News Ontology Event Model (NOEM) to represent 5W1H semantic elements of an event and relations among events. The extracted semantic event elements are first represented as RDF (Resource Description Framework) triples, and then automatically imported into NOEM as instances.

## III. PROPOSED SOLUTION

Events are dynamic data structures that play key role in understanding real world eventuality, which are basically driven by the four Ws' (what, who, when, and where). This natural evolution of questions is a typical example of what one might ask about an event. Without any one of these questions, the news article telling about event would fall level and quickly become less attractive. Event involves complicated relations between people, places, actions and objects. Event centered modeling captures the dynamic aspects of an event and participating entities. According to assumptions made in this work and to capture and represent event facts, researcher defined event using four dimensions.

[Event E]: Defined an event E as a real world phenomenon that occurred at specific time T, involving participant P, and tied to a location L that published or posted to one of the public media available on the Web (News broadcast, social media).

$$\boldsymbol{E} = (T, P, L, K) \quad (1)$$

Where,

*T*: temporal information describes the time *when* an event occurs.

*P*: entities participated or intended audience (at real time or after), to describe *who* was involved,

*L*: spatial information (location or place) to describe the place *where* it took place,

*K*: textual description of the event, describe *what* is published,

The proposed event modeling architecture consists of five major components: preprocessing, event trigger identification, event semantic elements extraction, classification and event representation. Figure 1 shows the architecture of event extraction and representation model from news articles.

Preprocessing: is initial task for event extraction process. The main task of preprocessing module is to make the news article ready/soft for event extraction processes. It accepts unstructured raw natural language texts as input for processing. In preprocessing language specific subcomponents like tokenization and normalization are performed. The incoming news articles are plain text and to find where do sentence start and end, identify words, whitespace and punctuation marks. Sentence and word tokenization is performed using Amharic language punctuation marks and white space. Another task of preprocessing is normalization, mapping the recognized expression to their specific standard format or measures. Amharic language specific normalization like character and temporal expression normalization is performed here. Character normalization is responsible to normalize different Amharic characters having same pronunciation and meaning into single character. Character normalizer iterates over news article to find redundant characters and replace with norm character. To resolve the ambiguity of temporal expressions, a rule based temporal expression normalizer implemeted. The result of the preprocessing module is segmented news article into sentences and tokens, normalized (character and temporal expression), that makes news article ready for other components.

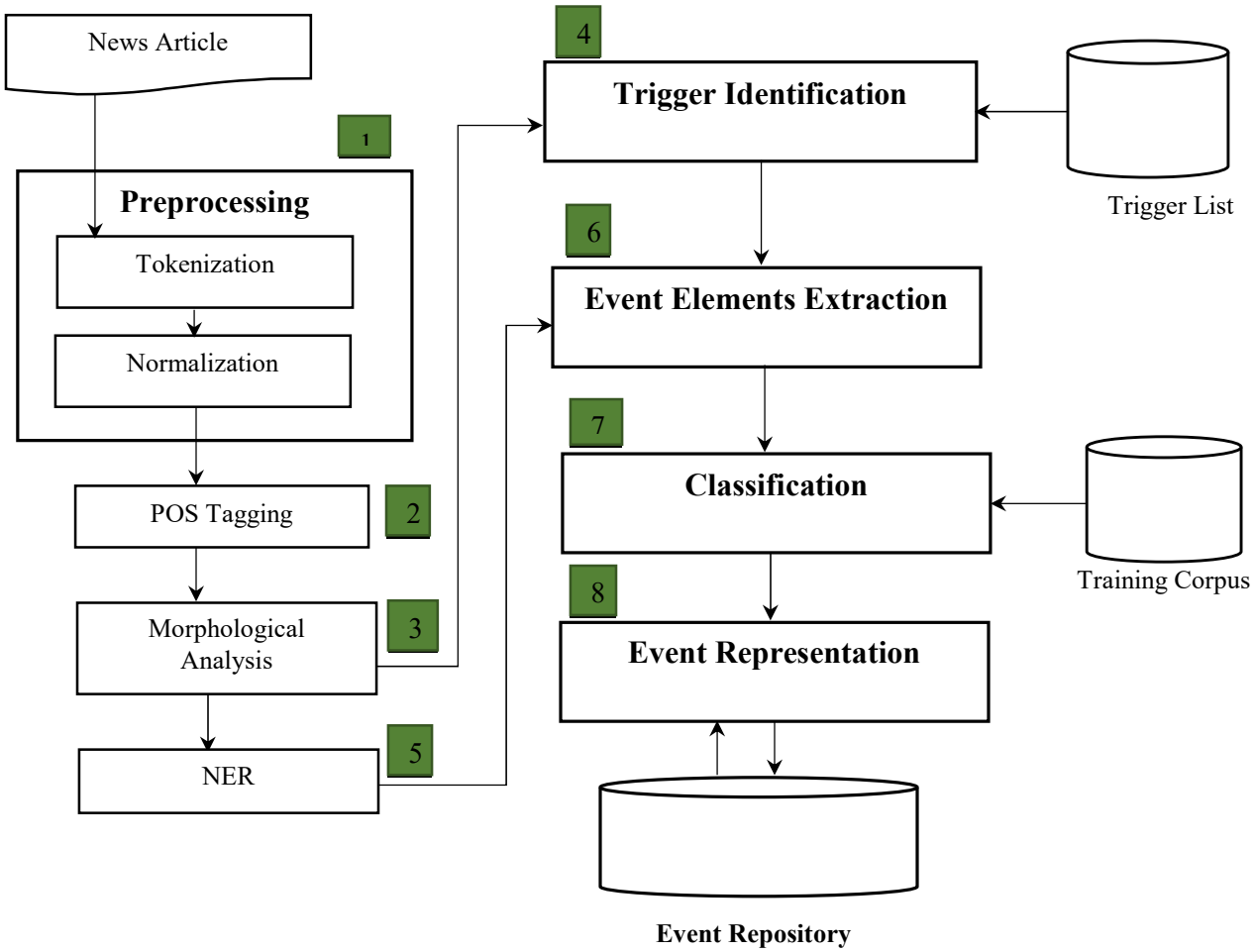


Figure 1: The Architecture of Event Extraction and Representation Model from News Articles

After news article is tokenized the next task is tagging the part of speech of determined phrases or tokens. The POS tagger receives tokenized news article as input and assigns POS tags to the words of that sentence that is output the lexical category of tokens such as nouns, verbs, adjectives, etc. Part of speech is used for filtering candidate triggers in event trigger identification. Usually events are triggered by verbs and nouns. To reduce the variation among event triggers and the sparseness for lexical representation of the text morphological analyzer is used. Morphological analysis is applied to list of candidate event trigger extracted from news article and trigger list. To handle morphological variety of tokens manually constructed Amharic morphological analysis is used. It accepts words and reduces to its root word.

Trigger Identification: Event trigger is a word or phrase starting an event, which is an important feature for recognizing the mention of event. Such words are found through extensive analyze of textual features in order to select the best contributors to the task of event identification. For this task collected event trigger words or phrase from the training corpus containing news article of different domains on the ground truth positive instances and sample list of identified event trigger words. Every word in news article can't be candidate event trigger, event trigger identification algorithm filter and use only possible word classes. Possible word classes are: verbs and nouns that describe events as they take place are considered typical features. To select most informative triggers that can tell the occurrence of event part of speech tagger, morphological analysis and a set of features like term of frequency value and position of candidate trigger is used.

Algorithm 1: Event Trigger Identification Algorithm

```
Input: Preprocessed news Article(A)

Output: event trigger

Begin
POST//Part-of-Speech Tagger
MA//Morphological Analysis
For every term in A

 If POS(term) in A == VB or NN
    Reduce term into its root form using MA
    Compare root word of candidate trigger with root of
    trigger list
    If trigger term match
       Compute candidate trigger position
       Compute TF(term)
       Compare the values
    The candidate trigger containing with highest
    value is recognized as key event trigger
   End if
  End if
End for

End
```

As indicated in Algorithm 1, trigger list and other local features indicated in algorithm to identify event triggers form news article. The true triggers head words are either verbs or nouns. Therefore, the algorithm considers only those words whose part of speech tags belong to these word classes as candidate triggers. Specifically, when applied

to candidate token ct and trigger list tl, S(ct, cl), checks whether ct is similar to any of the triggers in the trigger list tl. This is done using string similarity measure between a given ct and the trigger in trigger list tl. Other local features like term of frequency and position of candidate trigger have its own contribution to trigger identification. Term of frequency is the number of candidate trigger appears in news article. Candidate trigger with higher term frequency is more valuable than others. Position attribute is position where a candidate trigger first appears in the news article. A candidate trigger that occurs at the beginning of the news article is often more valuable than a candidate triggers that occurs at the end of that article.

Event Elements Extraction: Event elements are arguments that tell entities involved in the course of event. Event elements extraction component is responsible to extract basic information that is event elements like extent, participant, temporal and spatial elements. Each element can be described in very different ways, and the values are captured by processing the news article. Typically, event elements are described in media using the facets of what, who, where, and when of news articles. Extracting event elements that tells the entity discussed in news article is more important in event extraction. A number of entities can be mentioned in news article, but only valuable elements those entities and values that occur within the scope of the event is extracted. To select event elements named entity recognizer and other local features like potential trigger and path from the extent to head word of the trigger. Named entities are one of the most often extracted types of tokens during information extraction. To extract event elements from news article manually tagged news named entities is used. The four event semantic elements are identified by using named entity recognizer. Event elements are categorized into different class of named entity like people, organization, place, time etc. Table 1 shows event elements and their corresponding entity classes is presented.

Table 1: Event Elements and Possible Entity Classes

| **Event Element** | **Entity type** | **Description** |
|---|---|---|
| Temporal elements | TIME | Describe time *when* (date, year, instant, etc.) the eventuality takes place. |
| Spatial elements | LOC GPE | Describe place *where* an event happens. |
| Participant(s) elements | PER ORG | Describe person or organization *who* participated during the course of event. |
| Event extent (title) | -- | Describe *what* is mentioned in news article. |

Event extent is a short textual description of an event. Event extent extraction corresponds to selecting sentence that best expresses the idea or mention of event. Event extents are extracted using the result of measurement of the representativeness of the sentence. To determine the ranking of the sentences of a news article, a relative significance score is determined and the top ranked is chosen as event extent. Several features are used for ranking the sentences in the news article. Combination of various features of the sentence like position of the sentence in the news article, length of the sentence, TF*IDF score of the sentence, number of named entities and presence of trigger words in the sentence is used to extract event extent.

The temporal element is one of the key elements of an event which answers when event take place and its description can be a single day or it can be a time period with a starting and ending time. Temporal element extraction is responsible to extract date related expression from recognized named entities and select most relevant feature that relate to the occurrence of event. Most importantly, there are two temporal coordinates of interest in event presentation, real world time and media time. The real world time refers to time when an event takes place in the physical world, while media time refers time when event published over media, that could be immediate, or after period of time. As described in Table 1, word class of temporal element is TIME. A number of time references can be extracted from a news article, but the one closest to event extent is more important. Some events don't contain any date references or not reliable, in such cases media time is considered.

Space or location is another key media variable to interpret events. It answer question where an event takes place. Similar to temporal expression, the description of location can also take many forms absolute, relative or approximate. Location can be a city, area or a whole country. In order to determine event location named entity recognizer is used. Entity with class of location (LOC) and Geopolitical Entity (GPE) is a candidate for spatial reference. A number of entities with these classes can be found, but element within event extent or closer gets higher probability. Event participants are entities that are involved during the course of an event. Event participant has close relation to event extent that is sentence from where event trigger is extracted. Most common candidate of participant elements are person (PER) and organization (ORG) classes of named entity recognizer.

Classification: Event extraction is typically classifier-based and often uses training data set. Beside this, machine learning approach solves the problem of domain independence and considers event extraction as a classification problem, which focuses on the feature selection and classifier design. The goal of classifier is to distinguish between event and non-event data, filter event related data sources from nonevent using deterministic feature values of input news article and training corpus. In order to differentiate between triggers that represent event (positive instances) and those that do not (negative instances), or to choose the correct class label for every event candidate that represents an event machine learning approach is used. Machine learning algorithm integrated in our architecture makes use of various linguistic features. A feature corresponds to a specific property associated with an instance, and makes the connection between the instances that are observed and the categories that are to be predicted. To identify the class of candidate event into event and non-event, supervised machine learning algorithm called Maximum Entropy Classifier is applied. Maximum entropy classifier is used to model the known conditions and ignores unknown conditions.

Event Representation: The level of event representation has its own impact on human understanding or interpretation of events. Here in this work event structure is deeply covered for understanding the actual event information. Event representation is quite specific and includes not only actions, but also participants, time, and place. The main responsibility of this component is to map the extracted event facts to the designed event representation model. Ontology based event representation model is used. Ontology based event representation provide the basis for event representation that is semantically structured representation of events. Here ontology as a main mechanism to represent event information and construct event repository. Event centered modeling captures the dynamic aspects of an event.

## IV. EVALUATION AND EXPERIMENTAL RESULTS

Evaluation covers evaluating the performance of the proposed model through analysis of main components like key event identification, event elements extraction, classification and representation. A manual tagging of Amharic news articles has been made by seven annotators with the purpose of comparing it with automatically extracted by the system. Manual tagging is done by randomly selected four Computer Science postgraduate students and other three Information Science students. The task of manual tagging includes trigger identification and event elements extraction. Manual tagging is carefully done by distributing up to 14 news articles for a single annotator. To compute the performance measure of our system, most widely used information extraction measures like recall, precision and F-measure is used.

To identify the mention of event in news article, researcher used the guideline of ACE [12] and manually collected 220 events trigger words and phrases from various news domains. The evaluation is made with the parameters that compare the manually tagged and the one identified by the system. As previously discussed manual tagging include trigger identification, extent identification and event elements extraction. The basic idea here is, if the identified trigger contains the human tagged trigger, mark the identification as correct, otherwise as wrong.

Evaluation result shows, the trigger identification obtain precision (67.1%). The experimental result also shows event triggers are moe important to detect the mention of event and identification of event extent. Event identification use different features for event trigger identification, as a result trigger found at the beginning of news article is more important than trigger located at other position. The result also shows that, algorithm tends to be affected by the length of news article and number of candidate trigger identified in news article. The result of event elements extraction obtains precision values for event extent extraction (68.2%), temporal element (77.6%), spatial element (70.6%) and participant (60.0%). Most often news writer uses a common temporal expression that is why higher precision than other elements. In addition to this the experimental result shows that, the performance of event element extraction is directly influenced by the performance of named entity recognizer. Enhancing the named entity recognizer will improve the performance of event elements extraction. The aggregate performance of event elements extractor obtains precision (69.1), recall (79.4%) and F-Measure (73.9%).

*Table 2: Comparison of extracted event element with manually annotated articles*

| Event Elements | No. of Similar Elements | No. of Dissimilar Elements | No. of Empty | P | R | FM |
|---|---|---|---|---|---|---|
| Event extent | 58 | 5 | 22 | 68.2 | 80.6 | 73.9 |
| Temporal Elements | 66 | 5 | 14 | 77.6 | 91.7 | 84.1 |
| Spatial Elements | 60 | 12 | 13 | 70.6 | 83.3 | 76.4 |
| Participant Elements | 51 | 8 | 26 | 60.0 | 72.9 | 65.8 |
| Average | 235 | 30 | 75 | 69.1 | 79.4 | 73.9 |

Evaluation of the classifier using dataset consisting of 1225 event related news article collected from Fanabc news archive. Then reshuffled the collected news articles, the top 980 news articles used as training data and the remaining 245 news articles are used to test the performance of the classifier. After training the Maximum Entropy classifier (MaxentClassifier) using training data, then evaluate the performance of the trained classification algorithm using unseen test data. The accuracy measure is the percentage of inputs in the test data that the classifier correctly classified. The result shows, machine learning classifier module correctly classifies (72.0%) of the events. The performance of classifier depends on the performance of feature extraction. Thus, with lower NLP tools our classier obtain good result. The result also shows that event trigger is more important than other elements, because it signals the mention of event.

Janez Brank et al. [13] classified ontology evaluation methods into four categories: comparing the ontology to a golden standard, using the ontology in an application and evaluating the results, comparisons with a source of data about the domain to be covered by the ontology, and evaluation by humans who try to assess how well the ontology meets a set of predefined criteria, standards, requirements. Event representation model is evaluated with respect to the requirements, and event dimensions considered in this work. The designed event representation model provides a basic vocabulary for semantic representation of event elements identified in news article. For this classes and properties are carefully selected to guarantee the model compactness as well as to supply rich semantics. It captures temporal, spatial, informational aspect of events and media information. Entities like people and objects that participate in an event are described. In addition to this, the designed event model is able to cover information of events in two levels, which are event information (time, place, entity and relation) and event media information. At same time our model has more compact design with only few classes and properties, it is suitable for modeling Amharic news articles.

## V. CONCLUSION AND FUTURE WORKS

Events are central aspect for human to perceive the real world eventuality. Event scenario is driven by the 4W's (what, when, where, and who) that is made up of event trigger and event elements. Seminal events are published in news article or social media like any mundane events. In this study, the researcher proposed an event extraction and representation model from news articles. Event modeling task involves seminal event identification, semantic event elements extraction and representation. Event trigger identification and event elements extraction is done through extensive language specific linguistic analyze of unstructured news article. To identify the mention of event, 220 event triggering words and phrases collected from various news domains. Named entity recognizer and other local features are used to extract event elements. Maximum Entropy classifier algorithm is used to classify news articles telling about event into real world event and nonevent. The classifier is trained and tested by 1225 event related

news article collected from Fanabc Amharic news archive. The identified events are semantically represented and stored in event knowledge base which can be used for various event level semantic applications.
Evaluation of the proposed work is carried out by comparing manually tagged news article with the developed system. With lower NLP tools, our evaluation of event extraction model shows good performance. The event trigger identifier module obtain precision (67.1%) of event correctly which contributes to the better event element extraction and classification. The event elements extractor component shows greater obtaining precision (69.1%) while event classification module classify about (72%) of event correctly. In general, the proposed solution has shown a promising result when compared with highly resourced language event extraction systems.
Contribution of the Work: The main contributions of this research work are summarized as:

- Event mention identification in news articles,
- Event elements extraction, concept of 4W semantic elements in news articles,
- Classification of published news article into events and nonevent.

Future works: Event extraction from free unstructured natural language is a very complex task, which consumes more time, and can be enhanced by the performance of different NLP tools. Hence, there are a number of gaps for improvement and modification for event extraction and representation from news articles. Here are some of the recommendations researcher propose for future work.

- Incorporating Semantic Role Labeling: incorporating SRL will bring additional relevant features that can be used to identify the role of participating entities. In addition to this additional event dimension like *why* and *how* can be extracted by reasoning.
- Enhancing NLP tools: for this work minimal NLP tools is used. The performance of NLP tools like POS tagger, named entity recognizer and morphological analysis have significant impact on extraction process. Thus using enhanced NLP tools can improve the performance of the system.

## REFERENCES

[1] Wang Wei and Zhao Dongyan, "Chinese News Event 5W1H Semantic Elements Extraction for Event Ontology Population", *International World Wide Web Conference Committee (IW3C2), WWW 2012 Companion*, April 16–20, 2012, Lyon, France, 2012.
[2] Miller G., Beckwith R., Fellbaum C., Gross D. and Miller, K., "Introduction to WordNet: An online lexical database", *International Journal of Lexicography 3(4),* 235–312, 1990.
[3] Fazli Can, Seyit Kocberber, Ozgur Baglioglu Suleyman Kardas, H. Cagdas Ocalan and Erkan Uyar, "New Event Detection and Topic Tracking in Turkish", J*ournal of the American Society for Information Science and Technology,* October 1, 2009.
[4] Wei Wang and Dongyan Zhao, "Ontology-Based Event Modeling for Semantic Understanding of Chinese News Story", *NLPCC 2012*, CCIS 333, pp. 58–68, 2012.
[5] Nasser Alsaedi and Pete Burnap,"Feature Extraction and Analysis for Identifying Disruptive Events from Social Media", *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Cardiff School of Computer Science & Informatics Cardiff University, Cardiff, UK, 2015.
[6] Zheng Chen and Heng Ji, **"**Language Specific Issue and Feature Exploration in Chinese Event Extraction", *365 Fifth Avenue*, New York, NY 10016, USA.
[7] Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum, "Yago: A Core of Semantic Knowledge Unifying WordNet and Wikipedia", *In Proceedings of the 16th international conference on World Wide Web*, pp. 697–706 ACM, 2007.
[8] Daniel Hienert and Francesco Luciano, "Extraction of Historical Events from Wikipedia", *ESWC 2012 Satellite Events*, LNCS 7540, pp. 16–28, 2015.
[9] Axel Magnuson, Vijay Dialan and Deepa Mallela,"Event Recommendation using Twitter Activity", RecSys'15, *ACM*, 978-1-4503-3692-5/15/09, September 16–20, 2015, Vienna, Austria, 2015.
[10] Hila Becker, Dan Iter, Mor Naaman and Luis Gravano, "Identifying Content for Planned Events Across Social Media Sites", WSDM'12, ACM 978-1-4503-0747-5/12/02, Seattle, Washington, USA, 2012.
[11] Jakub Piskorski, Hristo Tanev, and Pinar Oezden Wennerberg, "Extracting Violent Events from On-Line News for Ontology Population", *BIS 2007*, LNCS 4439, pp. 287–300, 2007.
[12] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The Automatic Content Extraction (ACE) program-tasks, data, and evaluation", In LREC, 2004.
[13] Janez Brank, Marko Grobelnik and Dunja Mladenić, "A Survey of Ontology Evaluation Techniques", In: *8th International Multi-Conference Information Society (IS 2005),* pp. 166 170, 2005.
[14] Farzind Aratefeh and Wael Khreich, "A Survey of Techniques for Event Detection in Twitter", *An International Journal of Computational Intelligence*, Volume 0, Number 0, 2013