International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:12, 2012

Event Template Generation for News Articles

A. Kowcika, E. Umamaheswari, and T.V. Geetha

***Abstract*—**In this paper we focus on event extraction from Tamil news article. This system utilizes a scoring scheme for extracting and grouping event-specific sentences. Using this scoring scheme event-specific clustering is performed for multiple documents. Events are extracted from each document using a scoring scheme based on feature score and condition score. Similarly event specific sentences are clustered from multiple documents using this scoring scheme. The proposed system builds the Event Template based on user specified query. The templates are filled with event specific details like person, location and timeline extracted from the formed clusters. The proposed system applies these methodologies for Tamil news articles that have been enconverted into UNL graphs using a Tamil to UNL-enconverter. The main intention of this work is to generate an event based template.

***Keywords*—**Event Extraction, Score based Clustering, Segmentation, Template Generation.

I. Introduction

EVENT extraction is a particularly challenging category of Information extraction (IE). Information retrieval systems [1] are responsible to provide the information of about user's interest. Information Extraction systems rely on a set of extraction patterns that they use to retrieve relevant information from each document in the corpus. Event extraction involves the identification of instances of a particular type of event in free text, and the identification of the arguments of each such specific event. That is identifying who did what to whom, when, with what methods, where and why. There is now considerable literature on supervised and semi-supervised methods for event extraction. Most current event extraction systems rely on local information at the phrase or sentence level. However, this local context may be insufficient to resolve ambiguities in identifying events. Identifying an event and its participants and event attributes is quite difficult because a larger field of view is often needed to understand how facts are tied together in the article. Sometimes it is difficult task even for people to classify events from isolated sentences.

News articles typically discuss a new event in detail and mention relevant information about the people and person involved in that event. Still, a reader who may not have previously heard of one of these entities may want to see a biographical sketch. Alternately, reading an article describing an organization as having the "second highest revenues in its industry," may want to know what those revenues are exactly, and for what goods and services. Or in reading about an event that just occurred, may want to know about the earlier events that led up to it. In many cases one may find this information later in the article, in the background it provides for the event it covers. In focusing on details that are new or have changed, however, articles often leave out contextual information of this sort. The proposed system will analyze this information and extracts the temporal expressions.

A. Kowcika is with the Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai.TamilNadu-600025, India. (phone: +919944934634; e-mail: kowci.mars@gmail.com).

E. Umamaheswari is with the Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai.TamilNadu-600025, India.

T.V. Geetha is with the Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai.TamilNadu-600025, India.

II. Literature Survey

Event detection focuses on the automatic identification and classification of various event types from the given corpus. Event detection is treated as a sentence identification problem where the detection of the sentences associated with each event instance occurs is been described by M. Naughton et al [6]. There is a possibility that there exists the problem of identifying the mentions associated with each event, a problem which must be carried out by systems participating in ACEs VDR task, where this problem is not studied directly or evaluated in isolation within ACE. In order to develop methods for this form or indeed for any form of event detection, it is important to have a firm understanding of what is meant by an event and associate terms with it.

The simple breakdown of the task embodied by the system and the limited feature engineering for the machine learned classifiers; the performance is not too far from the level of the best systems at the 2005 ACE evaluation. David Ahn et al [9] describes an approach that is modular, and it has allowed to present several sets of works exploring the effect of different machine learning algorithms on the sub-tasks and exploring the effect of the different sub-tasks on the overall performance (as measured by ACE value). This is clearly a great deal of improvement. Improving anchor and argument identification will have the greatest impact on overall performance, and the experiments done. For anchor identification, taking one more step toward binary classification and training a binary classifier for each event type is a great deal.

The System of Shasha Liao et al [1] uses document-level statistical model for event trigger and argument (role) classification to achieve document level consistency within-event and cross-event. It shows the improvement in the performance of a sentence-level baseline event extraction system using document-level information. The model presented a simple two-stage recognition process; nonetheless, it has proven sufficient to yield substantial improvements in event recognition and event argument recognition. Richer models, such as those based on joint inference, may produce

even greater improvements.

The system proposed by Naushad UzZaman et al [2] extracts temporal information from raw text. This is a combination of deep semantic parsing with hand-coded extraction rules, Markov Logic Network classifiers and Conditional Random Filed classifiers. The system is compared and doing the same task on TimeBank corpus. The system outperforms the compared systems in event extraction, temporal expression relaxed match, temporal expression type and normalized value identification, event class identification and does equally well as compared systems in other event features and exact match temporal expressions.

The framework proposed by Kolikipogu et al [3], processes the time oriented information in legal text documents. It helps to find the multiple legal attempts if any information based on time event. Most of the temporal reasoning systems limit to analyze the past behavior for decision making systems. Those fail in finding multiple cases on the same timestamps. This attracted to propose a novel approach that addresses the timed event extraction and reasoning. This model starts with NLP prepossessing techniques followed by multiple knowledge bases, and temporal reasoning approaches. This system aims to support legal practitioner lawyers by providing temporal relationships for decision making. This model can be even implemented for Domain oriented Event Extraction and Reasoning.

As a subtask of information extraction, temporal information extraction aims to extract time expressions and temporal expressions from natural language text, and represent them in a structured knowledge framework. This area of research receives growing interest in NLP. It has been applied to question and answering, information extraction, text summarization, event extraction systems and temporal text understanding. The system proposed by Kam-Fai Wong et al [4] presents an overview of this research area. It also presents view on future research works. It forecast that it will become a challenging research topic in computational linguistics and artificial intelligence.

III. Methodology

A. UNL Enconversion

In this work we used the UNL graphs as the input. Universal Networking Language is an intermediate language that processes knowledge across language barriers. UNL captures the semantics of the natural language text by converting the terms present in the document to concepts. These concepts are connected to the other concept through UNL relations. There are 46 UNL relations like plf(Place From), plt(Place To), tmf(Time from), tmt(Time to) etc . This process of converting a natural language text to UNL document is known as Enconversion. The UNL document is normally represented as a graph where the nodes are concepts and edges are UNL relations. An example UNL graph is shown in Fig. 1. Example 1: The Bomb was exploded at the Cricket Stadium by 5.30PM.

The nodes of graph namely, “bomb (icl>weapon)”, “explode (icl>action)”, “Eden Garden Stadium (icl>place)”and “evening (icl>time)” represent the terms bomb, exploded, eden garden stadium and evening present in the example 1. The semantic constraints in the concepts, “icl>weapon”, “icl>action”, “icl>place” and “icl>time” denotes the context in which the concepts occur. The edges namely, “obj”, “plc” and “tim” indicates that, the concepts involved are object, place and time. From the above example, it is shown that the UNL inherits much semantic information from the natural language text and portrays in a language independent fashion. The proposed work uses Tamil language text documents, enconverted to UNL for event extraction which is described in the next section.

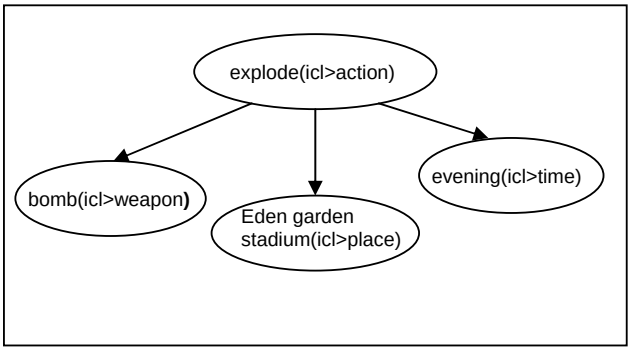

Fig. 1 UNL Graph

B. Reading Graphs

The UNL graphs contain the sentences whose contents are stored using multilist data structure. These multilists have to be processed where these processed graphs are given as the input for the proposed system.

C. Segmentation and Clustering

The UNL graphs are used as the inputs for the proposed system. Each sentence that has the following properties is considered as event specific sentence. The properties are that the semantic constraint in the UNL graph must contain icl>event or icl>activity and icl>person or icl>place and pos as dur. If these conditions are satisfied then those sentences are considered as event specific sentences. These sentences are extracted. Other non-event sentences are eliminated.

A new scoring scheme is introduced for identifying event specific sentences. Each sentence is checked with conditions and scores are added according to the similarity of the sentences. The conditions checked for identifying events are feature score and condition score. They are added up to get the probability value. The probability values between the sentences are obtained. The sentences with maximum probability value are grouped under that particular event. Multiple events will be obtained for each document.

The scoring scheme involves the following features for calculating the similarity between event specific sentences for grouping same events under same segment.

Feature score: Semantic constraint in the UNL graph if it contains icl>person the value added will be 0.2; icl>place the value added will be 0.2 and the pos must contain dur then the

value added will be 0.1.

Condition score: Semantic constraint in the UNL graph if it contains icl>event the term is considered as event; icl>act the value added will be 0.2 and pos as noun the value added will be 0.2 and the frequency is above 3 then the value added will be 0.1; icl>activity the value added will be 0.2 and pos as noun the value added will be 0.2 and the frequency is above 3 then the value added will be 0.1.

Similarity Score:

Similarity Score = Feature Score + Condition Score

The probability values between the sentences are obtained. The sentences with maximum probability value are grouped under that particular event. Multiple events will be obtained for each document. If the similarity score is above certain threshold then those sentences are grouped under specific event segment.

Event specific clustering is performed. The same scoring scheme is used for clustering event specific sentences. This is the inter-document clustering where events from multiple documents are clustered using the scoring scheme. The conditions checked for clustering events are feature score and condition score. They are added up to get the probability value.

The probability values between the segments are obtained. The segments with maximum probability value are grouped under that particular event clusters. Multiple events clusters will be obtained from multiple documents.

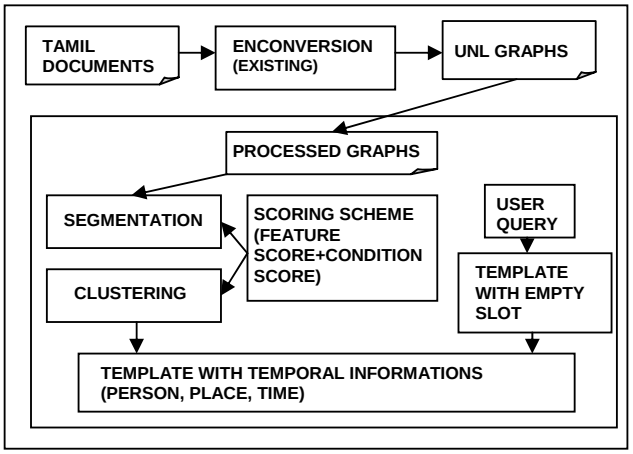

Fig. 2 UNL Graph

D. Template Generation

From the clusters obtained the temporal details are stored using hashmap data structure. The temporal details are placed where the event took place; person-who involved in that event; time-when the event took place. The template will fill the empty slots for temporal details with the user-specified query.

IV. Conclusion and Future Work

Event Specific sentences are extracted from the UNL Graph of sentences using the conditions. The conditions used for extracting event specific sentences can be modified that more conditions are added so that the efficiency of the proposed work can be further improved. The segmentation uses the scoring scheme in which the condition score and feature score can be further modified. The clustering also uses the same scoring scheme and this reduces the clusters formed as the number of input documents increases. Temporal information is extracted perfectly by the proposed system. The system can be enhanced by the following measures.

- Main events and sub-events can be identified from the given corpus using event frequency in the corpus and inverse document event frequency.
- Ranking can be performed based on the weightage of the particular event for the corpus documents using re-ranking algorithm.
- Event based indexing- Index will be generated for the user specified events.
- Temporal information based indexing – Index will be generated for the user specified events.

References

[1] Shasha Liao, Ralph Grishman, "Using Document Level Cross-Event Inference to Improve Event Extraction", in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

[2] Naushad UzZaman, James F. Allen, "Extracting Events and Temporal Expressions from Text", in Message Understanding Conference, 2010.

[3] Kolikipogu Ramakrishna, Vanitha Guda, Dr.B.Padmaja Rani, Vinaya Ch "A Novel Model For Timed Event Extraction And Temporal Reasoning in Legal Text Documents", in International Journal of Computer Science & Engineering Survey (IJCSES) Vol.2, No.1, Feb 2011.

[4] Kam-Fai Wong, Yunqing Xia, "An Overview of Temporal Information Extraction", in International Journal of Computer Processing of Oriental Languages Vol. 18, No. 2 (2005) 137–152, 2005.

[5] S. Sangeetha, R.S.Thakur, Michael Arock, "Domain Independent Event Extraction System Using Text Meaning Representation Adopted for Semantic Web", in International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) ISSN: 2150-7988 Vol.2 (2010), pp.252-261, 2010.

[6] M. Naughton, N. Stokes, J. Carthy. "Investigating Statistical Techniques for Sentence Level Event Classification", in Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, pp 617-624, 2008.

[7] Charles J. Fillmore, Srini Narayanan, Collin F. Baker, "What can linguistics contribute to event extraction?" in American Association for Artificial Intelligence, 2006.

[8] N. Stokes et al, "SeLeCT: A Lexical Cohesion Based News Story Segmentation System", in AI Communications ISSN 0921-7126, IOS Press, 2003.

[9] David Ahn, "The Stages of Event Extraction", in Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pages 1–8, Sydney, July 2006.

[10] M. Naughton, N. Kushmerick, J. Carthy, "Event Extraction from Heterogeneous News Sources", in American Association for Artificial Intelligence, 2006.

[11] Nancy McCracken, "Combining techniques for event extraction in summary reports", in Proceedings of the workshop on Event Extraction and Synthesis, AAAI 2006 conference, American Association for Artificial Intelligence, 2006.

[12] Zheng Chen, Heng Ji, "Graph-based Clustering for Computational Linguistics: A Survey", in Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, The City University of New York, 2010.