CHAPTER 2

LITERATURE SURVEY

In this chapter, we will first discuss the existing machine learning approaches to document processing tasks, which we have attempted to handle for implementing the news search engine. Subsequently, we will discuss how our search engine differs from the existing news search engine. Here, the news event extraction is the first task, and it is important to know the event arguments from the text corpus, for detecting specific knowledge about an event from the news text, and requires linking events to their associated time, places and participants. Moreover, the event extraction process can be applied for different domains, languages and genres of the documents. Different machine learning approaches have been attempted to extract events from the news text, but however many of them do not consider semantics for extracting event specific information. Therefore, in this chapter, we have done an analysis of news event extraction, and we have also discussed the related work for the other document processing tasks such as event clustering and temporal expression identification. Since we finally use our event extraction for news event search, the existing news event search methods have also been discussed in this chapter.

2.1 INTRODUCTION

Earlier work on event extraction is based on the use of handcrafted rules, which is time consuming to write for every event extraction task. An event cannot be predicted only with the limited set of rules; it can come in any
form. In order to extract events in an efficient way, considerable amounts of machine learning research work have been carried out using supervised, unsupervised and semi-supervised event extraction. Supervised approaches perform well and are suitable for applications that have the predefined knowledge. However, annotating sufficient information for the training process requires manual effort, which is time consuming. Unsupervised approaches do not require predefined knowledge, but it requires a very large corpus to get accurate results. Moreover, the amount of time required to reach convergence is large. Hence, additional effort is required to improve the performance of the system. Semi supervised approaches can reduce the annotation effort, and by using limited annotated seeds to learn and extract event based information converge time can be limited. Searching events requires handling event time, place and participants. In this context, the tackling of temporal expressions becomes important. Event clustering helps to group the specific event in each cluster for event indexing. The subsequent sections discussed the above mentioned text processing tasks such as event extraction, temporal expression identification, event clustering, which is required for the news event search. Finally, we have discussed the existing event search system for the news text.

2.2 EVENT EXTRACTION

Different approaches have been attempted to extract events. The most common among them are rule based approaches, machine learning approaches and statistical based approaches. The machine learning approaches can be classified into supervised, unsupervised and semi-supervised.

Earlier work on the news event extraction (Andersen et al 1992, Rau & Jacobs 1998) used rule based approach for template based pattern matching technique. These systems require predefined rules to specify the
domain specific event templates, and do not analyze the document for the syntactic or semantic based information; they have analyzed only words of the document for the frequently occurring patterns for filling the domain specific templates. Faiz et al (2002) also used a rule based system; by additionally considering the lexical and morphological analysis to extract the patterns, that have a similar morphological suffix and lexical information.

The above said rule based approaches can be used only for a specific domain, and a specific language. Writing rules for all possible event facts require a lot of effort, and even then may not cover all the event facts.

In order to adopt rule based systems for the different domains, Malik et al (2011) proposed an approach that combines both statistical and rule based approaches that use annotated training data to extract facts from the document. Compared to the previous approaches, this approach does not use domain specific rules; instead it uses a statistical classifier such as Singular Value Decomposition (SVD) and Feature weight classifier (FWC) to automatically learn patterns that fill the event templates.

All the above said approaches use only the hand crafted rules or learn patterns to fill event templates written for a specific purpose, but do not utilize the knowledge based systems to provide the event semantics. Some rule based approaches use semantic knowledge to improve the accuracy of the event extraction system. Borsje et al (2010) has developed a rule based event extraction system, which uses a word group lookup table, and a Morphological analyzer and a POS tagger, semantics in the form of ontology based gazetteer list. In their approach, the domain ontology contains information about the NASDAQ-100 companies, extracted from Yahoo! Finance. The event extraction was based on the GATE (General Architecture for Text Engineering) Rule Transducer component, which uses JAPE (Cunningham et al 2011) for manually defining the patterns. Since, this
approach has been developed for the financial news events, additional effort is required to build ontology for other domains. Therefore, Statistical based event modeling has been introduced by Wang et al (2012), that covers general event concepts and relations to define events. They have defined ontology based concepts, to capture the temporal, spatial, experiential, structural and causal aspect of any event. Solovyev et al (2013) proposed a semantic based approach for news event extraction that allows automatic extraction of business events using domain-specific dictionaries, where elaborated rules are conducted with the help of OntosMiner (Efimenko et al 2004) and RCO (RCO 2011). OntosMiner and RCO are the two semantic based information extraction systems that have been developed for Russian language.

Though, the above approaches cover general event relations and concepts for handling news events, some applications need to specifically analyze an event, where the updation of the new concepts and relations, is needed to handle events such as Natural disaster, Economy, Election, Sports, etc. However, building a detailed ontology (Sureshkumar et al 2013) for this type of events leads to difficulties in classifying the entities to a particular class. Machine learning approaches have been used to enhance event extraction in order to automatically learn event patterns without much manual effort. Machine learning helps in predicting hidden information, based on the known properties learned from the training data. These machine learning approaches have been classified into supervised, unsupervised and semi-supervised.

Most of the supervised approaches used Propbank (Palmer et al 2005), FrameNet (Baker et al 1998) or VerbNet (Kipper et al 2000) classes to extract event facts from the documents. Since PropBank and VerbNet convey the meanings of the words in a broad sense; the extracted events may not give the exact correct sense. Similarly, FrameNet (Johnson & Fillmore 2000;
Fillmore et al (2003) have also been used to extract event facts (Chambers et al 2011) from news text. The limitation in using all the above said semantic frames, are different types of multiple frames may be chosen for polysemous words, which may hinder the performance. Moreover, these semantic frames are not readily available for the most of the languages. Supervised approaches require predefined knowledge and the knowledge base systems such as FrameNet, VerbNet and ProbBank work for certain domains and languages only. In order to handle event extraction in a domain and language independent way unsupervised approaches have been attempted.

Unsupervised approaches have been attempted for different applications that depend on the statistical structure of the input, and provide only the general information for a given problem. Unsupervised approaches cannot be specialized to give answer for specific knowledge, since it is based on the probability distribution of the words that appear in the corpus, and requires large corpus to get better result and moreover cannot be easily adapted to different domains. Chambers et al (2011) introduced an unsupervised narrative event chain detection method based on the common protagonists. A distributional score was computed to find how often the two event verbs share the grammatical arguments (using point wise mutual information) to create the pair wise relation between them. In their work, Verb relations and their semantic roles were jointly learned automatically without predefined knowledge. Similarly, Ji & Chen (2008) introduced an unsupervised event extraction method, which used the existing information retrieval system (Strohman et al 2005) to identify the topically related clusters to group the related events, and also used HMM based classifier to identify events and nonevent sentences. They have introduced ‘centroid arguments’ to find the most frequently occurring entities for an event. Instead of focusing on the verb arguments, this approach considers the entities as centroid to extract
the event arguments. Ji & Chen (2008) use the most frequently occurring entities as centroid to find the link between the events.

A Latent Dirichlet allocation (LDA) based probabilistic model (Bejan et al 2008) has also been proposed to automatically learn the distribution of the event terms across the document. They have tackled 6 new event relations such as Sub event, Reason (Cause and effect), Purpose (Intension), Enablement (Event A allows an Event B to happen- not necessarily be a cause of an event), Precedence (sequential ordering of event) and Related events. Similarly, LDA based template learning for news event has been developed by Chambers et al (2011). They focused mainly on the automatic learning of the domain specific template structure using LDA. They have followed the Message Understanding Conference Series (MUC-4) event template structure to learn the event templates. The probabilistic based LDA model has been introduced by Li et al (2012). They have also used entity-driven topic modeling approach to extract the topically related entities, since entities and their arguments are more important to extract the events. All the previous approaches first find the centroid for an event, and then the associated event arguments related to the event centroid. The event centroid can be an event/ entities. Semi supervised approaches usually starts with a small set of known examples, and iteratively learn patterns to extract the informative patterns from the news text. Therefore, for a large application which requires extracting specific information with less computational effort, requires semi supervised approach to automatically extract the event facts.

Xu et al (2002) have proposed an Automatic Event semi supervised event extraction Domain Adoptive Relation Extraction (DARE) method that uses a bootstrapping approach. The DARE system automatically learns rules for any relations or events by providing a small set of linguistically annotated documents and some initial examples of the relations. This approach has been
applied for opinion mining and event extraction from German text. Rules are learnt based on the grammatical structure of the sentences, which are not sufficient to find complex relationships between one event and another. Similarly, cluster centric semi-supervised bootstrapping approaches (Piskorski et al 2008) have also been introduced to find similarity between events across documents. They have computed only the term level similarity between the two patterns. Piskorski et al (2008) uses Hidden Markov Model (HMM) based clustering algorithm to group the events. They use a lightweight manually tagged linguistic knowledge base, to learn the new pattern rules. Piskorski et al (2008) has used only on the lexical similarity between the two patterns which may lead to wrong interpretation. In order to obtain good patterns for complex events, Bootstrapping Events and Relations (BEAR) (Liu & Strzalkowski 2012) has described another semi-supervised bootstrapping approach that recognizes complex event semantics, using multi-pass bootstrapping. By considering context sensitive words as triggers, this approach can give meaningful results to the user. Huang & Riloff (2012) proposed similar approach that uses role identifying nouns as seed samples. However, considering the only event role identifying nouns as seed samples may not cover all the context of a given event noun. Therefore, in order to extract events semantically and to provide context based similarity, a Universal Networking Language (UNL) based bootstrapping approach has been proposed in this thesis. As explained in Chapter 1, UNL (UNDL 2005) is an Interlingua approach that represents the natural language sentences into directed graphs, in which it contains nodes and relations. The approach considers graph based semantic features that convey semantics at both word level and as well as sentence level. The event semantics and context can also be expressed across sentences by analyzing the conceptual link between the concepts that appear in one sentence with the concepts that appear in other sentences.
There has been sustainable research in the semantic based information extraction for the past 10 years, due to the huge growth in the information available on the web. Web documents contain a large collection of information related to different event types such as business, criminal, natural disaster, political, etc.Extracting, such information requires domain-specific knowledge, and therefore depending on the corpus, and the nature of the languages different approaches have been used. Next section discusses the related work in the news event clustering.

2.3 EVENT CLUSTERING

In this thesis, our aim is to cluster specific events into a single cluster. Hence, a richer set of features with event based feature weight are required to obtain specific events in a single cluster. This depends on the level of context taken for representing an event. Context representation is an important aspect of any event context clustering approach. The context of an event can be expressed by a single word, sentence, passage, and even at the document level. Different approaches take different levels of context suitable for their applications. Since, our event clustering approach is based on the semantic based context clustering; we have discussed the context level document clustering for event mining.

Word level context has been taken by Maynard & Anmfiadou (1999) as features for the document clustering. In their work, WordNet has been used for determining the meaning of the surrounding words. Their clusters gave errors mainly due to incorrect chunking, syntactic and semantic tagging. Similarly, Murua et al (2008) proposed a Gaussian Mixture Document Clustering (GMDC) algorithm for news documents, based the likelihood of the data, using the Bayes rule. All the above said approaches follows the traditional Vector Space Model (VSM), to represent the high dimensional documents feature space, which is high dimensional. This may
consume more memory and time to cluster the documents. In order to reduce
the high dimensionality of the feature space, the features need to be chosen
appropriately the number of terms/concepts considered to represent the
document can be reduced. Hung & Wermter (2004) extends the traditional
VSM and has introduced three VSM namely, extended significance vector
model (ESVM), the Hypernym Significance Vector Model (HSVM) and the
Hybrid Vector Space Model (HyM). They have also used the WordNet
ontology to extract semantic knowledge, which can further improve the
cluster accuracy. The frequent item set has been used for clustering (Krishna
& Bhavani 2010) to reduce the high dimensional space. In their approach,
instead of a term based representation, a group of terms related to an item set,
has been considered for computing cluster similarity between the documents.
They have used an Apriori algorithm to find the frequent item set from the
document. They have achieved non overlapping partition across the cluster
results, which improved the accuracy of the cluster results. Similar to the
above mentioned approach, frequent concepts based document clustering
(Baghel & Dhir 2010) has been proposed to reduce the high dimensional
space. In their approach, they have taken the related words as features, rather
than the bag of words. These related words convey the same meaning that
were conveyed by the concepts. They have also used Apriori based method, to
find the frequent concept set across the documents to compute cluster
similarity. Recently, Bora et al (2012) has also introduced a frequent term,
and noun based clustering approach. They have considered only the news
headlines to cluster the results. The statistics of the terms that appear in the
news headlines are used to find frequency of occurrence of the terms across
the documents. Though, this approach reduces the time and memory required
for obtaining the cluster results, their results may not always be satisfactory.
The reason is, when only the headlines are considered for clustering, which
may not cover all the information related to an event and moreover noisy
headlines may occur, which will also hinder the accuracy of the cluster
results. All the above said approaches consider only word level context for clustering.

Event based clustering has been improved by increasing the context to sentence level (Naughton et al 2006), where similar sentences are grouped together to represent a single event. Wang et al (2008) computes sentence – sentence similarity to build the similarity matrix. Symmetric matrix factorization has been used to group sentences into clusters. PropBank based semantic annotation has been utilized to analyze sentence level semantics, which also helps to find semantic relationship between sentences. However, in case of complex sentences, the sentences may have two or more events, which results in incorrect clusters. In that case, a sentence simplification task is needed; this often requires language dependent parser. Cai et al (2010) has also proposed a sentence based clustering approach that finds sentence level similarity based on the sentence mixture model using Expectation Maximization (EM) algorithm. The sentence mixture model was calculated, based on the distribution of the terms that appear in the sentence. Recently, Jeyalakshmi et al (2014) has introduced a fuzzy clustering approach that finds pattern based similarity between sentences. Their approach acts on the relational data, and overcomes the time complexity of the existing fuzzy clustering approach. A detailed analysis on the issues in sentence level clustering has been carried out by Sathishkumar et al (2013). They have stated that sentence level clustering may not give good result for short sentences, and most applications require finding more information by analyzing other sentences.

All the above approaches cluster the words or sentences that belong to a similar event into appropriate clusters, but do not differentiate between different events of similar nature or extract the attributes associated with the
event. Therefore, document level clustering has been attempted, which require processing detailed information.

Piskorski et al (2008) has also introduced an unsupervised cluster centric approach for extracting violent and natural disaster events. They have introduced a News cluster Event eXtraction Using language Structures (NEXUS) for automatically learning the event patterns to extract the events and their attributes. They have used the linguistic based linear patterns, which was specific to the natural disaster events. The use of linear patterns that depend on the order of the words/features, may not suitable for morphologically rich languages. Similarly, De Smet et al (2009) proposed a document-level event clustering based on the Probabilistic Topic Model such as Latent Dirichlet Allocation (LDA) (Blei et al 2003). This approach is based upon the probability distribution of the words that appear within the document together with all the entities. Their approach is suitable for clustering a large corpus with a diverse collection of words. A Latent semantic based approach for document clustering (Wei et al 2012) has been carried out, by taking both the title terms along with the sentences as features to represent the event context. The sentences, which match with the title terms, are considered as topic sentences, to represent the event context. They have considered only term level overlapping between the title terms with the document sentences to represent the document level event context. Usually, term level overlapping is minimal with the title text, and hence some relationships that appear in sentences that have no overlapping with title terms, may be missed. Therefore, in general we can conclude that the event context representation for a given document should consider the important sentences of the document, and the conceptual / semantic link between the concepts.

In order to provide semantic based cluster results, Choudhary et al (2003) introduced a UNL based semantic clustering that considers the
concepts as feature to represent the event context. From their observation, instead of computing term level similarity, the conceptual similarity can improve the cluster accuracy, however the links between the concepts are not considered to cluster the events which may further improve the cluster accuracy. The work described in this thesis extends the approach used by Choudhary et al (2003), by additionally considering the conceptual link between the concepts to represent the event context.

2.4 HANDLING OF TEMPORAL EXPRESSIONS

This section discusses the temporal expression identification and normalization from the natural language text. Most of the existing approaches, for the temporal expressions, use rule based approach. Mani & Wilson (2000) proposed a rule based approach with handcrafted and machine-learnt rules, to identify the temporal expressions from the news text. Yoon et al (2000) and Ferro et al (2004) both have designed a rule based system to handle the vague temporal expression, using lexical and syntactic based information from the text. The above said rule based approaches do not consider the semantics to identify and tag the temporal expressions.

In order to handle lengthy temporal expressions, Mazur & Dale (2007) used a rule based approach based on the fact that the functional dependency between the text tokens in the sentence along with the temporal trigger words, can determine the extent of the temporal expression. In order to handle semantics in rule based temporal expression identification, Detection And Normalization of Temporal Expressions (DANTE) (Mazur & Dale 2009) that uses semantic based inference and reasoning. However, it is difficult to handle lengthy temporal expressions, in which the scope of the temporal expressions is unpredictable. Since, the above said approaches uses only the lexical and syntactic features, it is suitable only for the structured
languages like English; however languages which are of agglutinative do not follow a standard structure to represent the temporal expressions. Moreover most of the rule based systems have difficulty in formulating rules for all cases, which is a bottle neck to adopt for different languages.

Some temporal expression identification uses annotated corpus, but was limited based on the availability of the annotated corpora for different languages. Allen et al (1983) first proposed the interval-based temporal expression algebra, to tag the temporal expressions in a standard way. Makkonen & Myka (2003) utilized the Allen’s temporal algebra, to tag the temporal expressions. They have utilized a syntactical parsing, and finite-state automata to know the temporal terms from the text. In their approach, the normalized temporal expressions across the document were compared with the other documents, to find the temporal relations such as Before, After, Continuous, Duration, Overlapping, Exact time.

TimeML (Pustejovsky et al 2003) is a specification language, for the annotation and normalizing of the temporal expressions of the natural language texts. These TimeML based annotations, help to tag the events and the temporal expressions, and to show the aspecual, temporal and subordinate link of the event with temporal expressions. Bittar et al (2011) utilized the TimeML based annotation for French texts, to tag events and temporal expressions. Similarly, Kolomiyets & Moens (2009) explored the differences in token and constituent based temporal identification, using an annotated corpus with the help of TimeBank. Both the approaches use only the grammatical aspects of the sentences, to tag the temporal expressions. However, it is difficult to process the temporal expressions semantically. A Semantic based temporal expression processing has been attempted by Llorens et al (2012). He used the lexical semantics and semantic roles, to
know the information about the temporal entities. In his work, the semantic roles have been identified with the help of ProbBank. He has compared his approach with the morpho-syntactic based approaches, and he has proved that his approach supports for different languages such as English, Spanish, Chinese, and Italian.

Automatic extraction of temporal expressions from text document can help to reduce time and efforts for tagging the temporal expressions. Different machine learning approaches have been attempted in supervised, unsupervised and semi supervised approaches.

Kolomiyets & Moens (2009) has attempted a supervised machine learning approaches, and compares the token based and the constituent based approaches, for automatically identifying the temporal expressions. This approach did not find the relation of the temporal expressions with the event. Kolya et al (2010) proposed a supervised machine learning approach, using the Conditional Random Fields (CRF), based on the TimeBank features. In their work, the temporal relation identification has been done as a pairwise classification problem, in which they find the relation of the time with an event within the single sentence, events with their Document Creation Time (DCT), and of the events with the other events in the adjacent sentences.

Unsupervised methods automatically learn temporal patterns without predefined knowledge. Mörchen et al (2007) proposed an unsupervised temporal expression handling and for the time points, they have used the sequential pattern mining algorithms for the time points with gaps in multivariate data. For univariate data and limited gaps suffix tree methods are more efficient. This is suitable for the documents which has similar nature. Poveda et al (2009) proposed a semi supervised bootstrapping approach for
handling temporal expressions. They have difficulties in handling complex, vague temporal expressions. The reason is that, their pattern representation considers only the sequence of tokens from the prefix and postfix of the temporal expressions with in a particular length. This is suitable only for explicit temporal expression. Some temporal expressions require the link between other temporal expressions to know the exact time of an event. In that case, the pattern representation should also consider the link between the temporal concepts in the sentence. Therefore, the work described in this thesis considers a semi supervised bootstrapping approach with semantic graph based features, which takes varying length of patterns, and also considers the semantic link between the temporal concepts across sentences.

2.5 NEWS EVENT SEARCH

News event search can be classified into time, person, place and multi-field based news event search. The time based news event search has been proposed by Allan et al (2001), Jin et al (2008), Berendt et al (2009), and De Sarma et al (2011). All their approaches focus only on the temporal aspect of the query terms. They have used only the term level features, the probability distribution of the terms, occurrence of terms for a given news topic as features. The dynamic relationship of the temporal expressions within the events has been attempted by Sarma et al (2011). They have used unsupervised clustering approach that finds time based similarity between the event entities. Most of the existing event tracking systems, gave importance to the temporal aspects rather than the person and place based event tracking. A multilingual person based software tool has been proposed by best et al (2006) that analyses large news collections and cluster related news, extract the named entities from the news clusters, and the name variants belongs to the same persons have been merged. They have used handcrafted rules to
tackle the name variants with the help of pronunciation dictionaries. A cross document temporal and spatial person based tracking system (Ji & Chen 2009) helps to track events in two dimensional, such as temporal person tracking and spatial person tracking. They have taken person entity as a centroid argument to find the confidence score of the other event arguments, which occurs more frequently. They have achieved cross-document person tracking. Kapp et al (2013) considers only person specific event exploration system and computes similarity based on the similarity between the participating person entities using Pointwise Mutual Information (PMI). They have constructed a person’s profile for each person detected in the corpus. For each person, they have analyzed the co-occurrence of the events within a specific window. These associations were indexed to track information about a specific person. In person and place based event tracking the spelling variations needs to be tackled by automatically learning the spelling variations and different names of a same person or place. Manually updating the dictionaries for all the spelling variations of the person and place based dictionaries requires high manual effort and even then may not cover all the names. In addition, anaphoric expressions need to be tackled to improve precision and recall of event search.

There is only limited work, which considers all the event entities for event search. The news event search, based on the event attributes is an important application of an event processing system. This has been attempted by Feng et al (2007) and Wunderwald et al (2011). Both of them follow only the rule based approach, and the semantic aspects were not tackled. The Ontology and the WordNet based semantics (Cybulska & Vossen 2011) has been used for event search. Compared to the rule based approach, Cybulska & Vossen (2011) incorporates semantics in event search. Another prototype
called NewsSync (Vydiswaran et al 2011), searched news stories, based on user preferences. They have provided a multi-field event search facility, however they store the event arguments in term level indexer, they do not consider the semantic relationship of the terms with other terms.

All the above-said approaches consider only co-occurring contexts and word level semantics while representing events. However, to build an effective event-based search engine a richer event model that should consider all the event arguments such as Time, Place and Person of an event are necessary. Moreover, the event semantics need to be considered at the document level. The reason is because events can have complex interactions with other events. Therefore, separating different events, and associating similar events require machine learning to automatically classify and cluster the news events.

2.6 CONCLUSION

This chapter discussed the existing work in event extraction, event clustering, temporal expression handling and event search, which deals with the discovering of event facts from the news documents. Supervised methods with a fixed set of features require predefined patterns, or labels to extract information from the document. Unsupervised approaches are suitable for the large corpus which does not have predefined knowledge, however the time required to reach convergence is high. In order to overcome the above problems, semi supervised approaches with limited knowledge have been proposed to automatically learn event patterns from the large corpus. This reduces the manpower and time required to reach convergence. The features used for different domains may vary depending upon the language, and domain. Fixed sets of features have been used to help to extract the information from the
structured languages in which the words appear in the specific order. However, it is not suitable for languages with agglutinative characteristics. In order to adopt the same approach for different purposes, the features should be chosen in a more generic manner, so that we can adopt the method with limited effort. Hence, a language and domain independent feature selection method can help the researchers to test the semi supervised approach for general domain.