Semantic Dissection of News Reports[1]

Sarwat Nizamani[1], Nasrullah Memon[2]

[1] University of Sindh, Pakistan, [2] Maersk McKinney Moller Institute,
University of Southern Denmark & Mehran University of Engineering and Technology,
PAKISTAN.

[1] sarwatnizamani@gmail.com

ABSTRACT

This paper presents the study of semantic analysis of news reports. The news is of immense importance for an analyst, who intends to analyze the law and order conditions of an area. In this paper we carried out semantic analysis of the new reports, which dissects the in a way that all the important pieces of information are highlighted. For instance, the main concept of the news, the entities involved, temporal and spatial information is extracted. The extracted information can help analyzer to comprehend the overall situation of an area. This method of analysis is efficient in a way that an analyst can give an instant look at news without going through the complete report, which save the analyst's time.

Keywords: Critical events, concepts, FBI News reports, information retrieval, semantic analysis

INTRODUCTION

This paper presents a study of semantically analyzing the News reports. With this method of analysis, critical events present in the News with all the involved entities, time of the event and location information is extracted efficiently. In the News archives, usually there are enormous amount of News reports which contain News in massive quantity. It is very intricate job to extract the information of interest from such a huge archive. One of the simplest techniques, which can be used to extract such information, is keyword based technique. Conversely, with the keyword approach, the context information may be ignored. The keyword approach treats each word equally, without any distinction of the context of the word in sentence. For instance, if there is a News about robbery incident that, 'Mr. X robbed Mr. Y', the keyword based approach will simply consider them as some X and some Y without distinguishing the suspect and victim. The keyword based approach will merely consider the occurrence of words in the News. This ignores the information that who caused the certain event and who was affected by the even, where and when event occurred? In this paper we performed the News analysis by employing the predicate argument structure [3]. This analysis proceeds by first identifying action words which are considered as predicates or main verb in the News sentence; next based on the information of that action word actors (entities) which are the arguments of that action, are identified, with the location and the time of the action information, if it is present in the News. Once, the entities are identified as arguments of the predicate, then depending on the context information of the entities involved, such as position in the sentence, type of the action word and word sense information of the action word, specific role (either entity caused the action or affected by the action) is given to that entity. The predicate is often a main verb in sentence and in English language a word/ verb may have more than one senses/ meaning. For each predicate, the

[1] The initial findings of this paper was presented at International Conference on Neural Information Processing (ICONIP2012) and published in LNCS.

ISSN: 2307-3721, e ISSN: 2307-3713
www.erint.savap.org.pk

number and type of the entities attached to the predicate vary from sense to sense. Thus, in this analysis we have also attached the sense information as a number. After the analysis is carried out the results can be observed by highlighting each critical part of the News. Hence, the contributions of the paper can be counted as:

1. An efficient method of News analysis
2. Highlights major concepts/ events
3. Distinguishing the victims or suspects of events
4. Highlighting temporal and location information

In the following sub-section, we describe the dataset, and then discuss the problem statement followed by outlining the paper structure.

FBI Dataset

The dataset is comprised of the part of FBI News reports[2]. These News reports are from period 2001 to 2011. The full dataset comprises of 4397 files belonging to various divisions of USA. We have taken a sample of 1056 files, which are national News in the dataset, each reporting the News of particular day. There is not a News report for each day but only the News if something special related to FBI happened is reported. The original dataset comprises of XML files and each XML file contains three elements, namely; title, date of publication and text. Initial preprocessing was required to extract text, and split text into sentences. This was implemented by using simple XML parser. Ones the text is available as sentence by sentence, and then News is ready for semantic analysis.

PROBLEM STATEMENT

We have an archive of News reports

Archive (News) = n_1, n_2...n_m

Problem of News analysis is to dissect the News reports from News archive in a way that main concepts of the News are extracted, such that following pieces of information are highlighted:

a. event/action/concept
b. entity who caused the event/action/concept
c. entity affected by event/ action/concept
d. location information of event
e. temporal information of event

Rest of the paper is organized as follows: Section 2 presents related work whereas Section 3 describes semantic analysis technique. Methodology is discussed in Section 4 and Concept extraction is demonstrated in Section5. The results are illustrated in Section 6 whereas conclusion with future directions is given in Section 7.

RELATED WORK

We present related work in this section by elaborating the use of the semantic analysis technique for various tasks. We do not include any comparative work to our article, because to best of our knowledge, we did not find any study on the dataset we used. We present the related work in two dimensions, namely; the studies employing SRL for different tasks; and the News analysis techniques.

[2] http://www.fbi.gov/news/stories

ISSN: 2307-3721, e ISSN: 2307-3713

We proceed by first discussing related work in connection SRL technique, Shen and Lapata [1] have investigated the use of semantic role labeling for question and answering problem. The authors experimentally show that the use of SRL improves the performance of question and answering task over the state of art. The authors [2] applied SRL for machine translation from morphologically poor to morphologically rich languages, such as from English to Greek and Czech. Authors claim that they achieved error reduction in verb conjugation and noun case agreement. The article [3] elaborates the use of predicate argument structure which is the basis of SRL for information extraction. Authors mention that information extraction task can remarkably be improved using the approach adopted. Kulick et al [4] have applied an approach to integrate biomedical text annotation using predicate argument structure of Propbank[3] and syntactic structure of Treebank[4]. The importance of Semantic Role Labeling can be realized from the fact that two consecutive CoNLL shared tasks [5] were dedicated to SRL in 2004 and 2005. The SRL task since then caught attention of NLP researcher and its importance was realized for the languages other than English. In this regard CoNLL shared task [9] in 2009 was dedicated to multilingual SRL.

Semantic roles have a great importance in language understanding [6]. SRL has a number of application including dialog understanding, question answering, machine translation, information extraction, dialogue understanding, word sense disambiguation and many more [6].

The authors [8] have emphasized on the use of SRL for Biomedical information extraction. The relation extraction is an important task for biomedical text extraction. SRL can efficiently extract these relations which are ignored by other statistical NLP tasks. The SRL considers all the relations for an event such as what, when, how, extent and so on [8].

As it can be realized from above discussion, SRL can be a very handy technique for analysis of the text. In this regard we conducted the study of News analysis using SRL. Therefore, we also discuss some related work to News analysis. For instance, the study [11] presents a news analysis method which identifies entities in the news and extracts spatio-temporal information of those entities. The visual analysis is then carried out which shows that how frequently certain entities appeared in the news articles in a specific time-window. Furthermore, the study extracts the co-occurrences of the entities in different news articles, which determines the relationships among the entities. We discuss the way we have adopted SRL technique in the study of News analysis in the following section.

SEMANTIC ANALYSIS TECHNIQUE

In order to semantically dissect the News reports SRL techniques has been employed. SRL analyzes the text at sentence level; each sentence is then examined against predicate argument structure. SRL approach responds a number of queries in a sentence. It analyzes the sentence by extracting pieces of information such that who did what to whom, why, how, when and where. For any News report, one expects all of these answers. For example if someone is interested to extract News regarding kidnap, SRL will return the query by providing information such as, who was kidnapped, who was the kidnapper, when and where kidnapping happened. This means that we can extract suspects, victims, instrument, location and temporal information about an event. SRL is based on predicate argument structure in which predicate mostly is a main verb of sentence that describes the major concept of the sentence. Each predicate is attached to certain roles which provide detailed information on

[3] Propbank is corpus containing annotations for predicate argument structure

[4] Treebank is a parsed corpus containing syntactic parsing in tree structure

ISSN: 2307-3721, e ISSN: 2307-3713

that action. We carried out the task of SRL on News reports using open source SRL tool[5]. The article [7] discusses the complete SRL model used in the tool.

Below we illustrate an example sentence from the dataset which is analyzed using proposed approach.

Following is an example sentence from the dataset, which is semantically, analyzed using the SRL technique

> "On Wednesday, November 18, 2009, at approximately 12:40 p.m. the Chase Bank, located at 16861 Bernardo Center Drive, San Diego, California was robbed by an unknown male."[10]

The sentence is taken from a News report, which is about a robbery incident of a Bank, by an unknown male located at 1686 Bernardo Center Drive, San Diego, on Wednesday, November 18, approximately at 12:40 p.m. We can observe by reading the News report that the event of robbery took place, the victim was Chase Bank, the suspect was an unknown male, location is 1686 Bernardo Center Drive, San Diego and time is Wednesday, November 18, at approximately 12:40 p.m. By the proposed analysis following pieces of information are returned:

- Who? unknown male
- What? robbery
- Whom? Chase Bank
- Where? 1686 Bernardo Center Drive, San Diego
- When? Wednesday, November 18, at approximately 12:40 p.m.

The methodology of the research is discussed in the following section.

METHODOLOGY

The research presented in the paper is based on analysis of text in order to extract critical information from News reports in comprehensible pieces. After the analysis has been performed on News text sentence, one can precisely conclude the report. The analysis can also help querying the reports against certain events/ actions, suspect, locations etc.

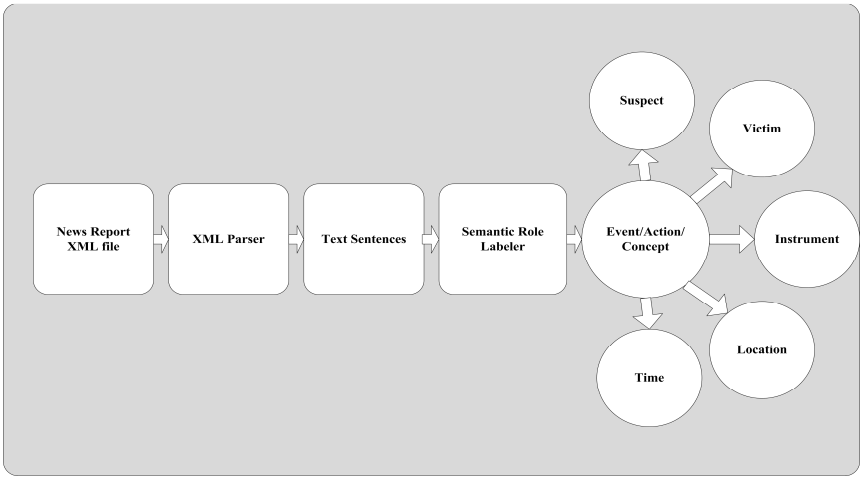

Figure 1. Semantic analysis process of FBI News reports

[5] http://code.google.com/p/mate-tools/

ISSN: 2307-3721, e ISSN: 2307-3713

The process of the News analysis is depicted in Figure 1. As it can be seen in the figure 1, initially we extracted the News in text, which were provided in XML format. News is analyzed sentence by sentence. From each sentence essential concepts are extracted with all the arguments. Initially a predicate is extracted, which is the core concept of the News. Then sense of the concept is disambiguated, afterwards entities related to that concept are identified and then these entities are assigned argument types such as A0, A1 and so on. The task of assigning roles to arguments of predicate makes the News analyses task robust as compared to keyword retrieval. Although the same can also be retrieved using keyword based approach but, the interesting idea of analyzing the News using proposed approach is to distinguish among the key event, the causer of the event and entity who affected by the event.

CONCEPT EXTRACTION FROM NEWS

As we previously mentioned that the predicates considered as the events described in News report. The events of interest which are extracted by semantic analysis of the part of the dataset are given below in the table with their senses. We extracted those predicates which may be of great interest of security informatics personal who want to analyze the law and order situation of specific area. We have generalized the arguments of the predicate, which are the related entities of a concept/ event.

Table 1. Prominent predicates found in the dataset

Predicate	*Causer*	*Affected*	*Other Argument*
assassinate.01	assassin agent (A0)	person assassinated(A1)	- - -
attack.01	Attacker (A0)	entity attacked (A1)	Attribute (A2)
bomb.01	bomb attacker (A0)	Entity attacked by bomb(A1)	- - -
commit.02	criminal	Entity affected (A2)	Crime(A1)
kidnap.01	Kidnapper(A0)	Person (s) kidnapped (A1)	- - -
kill.01	Killer(A0)	Person killed (A1)	Instrument
testify.01	Witness(A0)	Witness against (A2)	Evidence(A1)
surrender.01	Person surrendering (A0)	Surrendered to (A2)	Surrendered for (A1)
threaten.01	Agent making threat (A1)	Threat given to (A2)	Threat (A1)
mislead.01	Liar(A1)	Lied to(A2)	False statement (A1)
damage.01	Agent damager (A0)	Entity damaged (A1)	Instrument (A2)
murdering.01	Murderer (A0)	Person(s) murdered (A1)	Instrument (A2)
victimize.01	Victimizer (A0)	Victim(A1)	Grounds for victimization (A2)
violate.01	Violator (A0)	- - -	Rule violated(A1)

destroy.01	Destroyer (A0)	Entity destroyed (A1)	Instrument for destruction (A2)
detonate.01	Exploder (A0)	Exploded entity (A1)	- - -
offence.01	Offender (A0)	Offended (A1)	- - -
endanger.01	Exposer to dander (A0)	Entity in danger (A1)	- - -
evasion.01	Avoider (A0)	- - -	Thing avoider(A1)
warn.01	Entity giving warning (A0)	Warning giving to (A2)	Warning (A1)
prosecution.01	Prosecutor (A0)	Defendant (A1)	Law justifying case(A2)
mutilate.01	Mutilating agent (A0)	Entity mutilated destroyed (A1)	- - -
sabotage.01	Saboteur/ destroyer (A0)	Entity wrecked (A1)	Instrument (A2)
steal.01	Thief (A0)	Thing stolen from (A2)	Thing stolen (A1)
rob.01	Robber (A0)	Entity robbed (A1)	- - -
fake.01	Faker (A0)	Entity faked (A1)	- - -
hurt.01	Entity causing damage (A0)	Entity damaged (A1)	Instrument (A2)

In the table 1, we have only included three arguments specific to the predicate , some predicates have two, others have three specific arguments and some may have up to five, but the predicates we used in our study from the dataset have maximum three arguments. As it can be observed in the table above that the predicates extracted from the dataset show the actions which tell about the theme of News. Each action is related to certain entities, which are the causer or affected of that particular action. For example when concept of murdering is found in the News, then there is an entity murderer, who caused the murder which is also the suspect of murder and the entity murdered, who is affected by murdering action which is also the victim of murder, some other information can also be highlighted in the News such as instrument used for murdering, location information where murder took place and temporal information of the murder. It can be automatically answered from the proposed analysis method, that 'who murdered, whom, with what, when, where and why? if all such pieces of information are found in the News. In the table above A0 represents the agent argument or causer of an action, argument A1 is the patient or the entity affected by the action. In some predicates the affected of the action is represented as A2, because for these predicates A1 represent the certain action. For example in predicate threat, the entity who threatened is A0 or agent of threat, while threat itself is the argument A1 and the entity *threat given to* is A2. The other predicate such as *attack*, in which the *attacker* is the causer of attack which is the argument A0, the *entity attacked* is the A1 and the *instrument* used for attack is A2. The other arguments such as location and temporal are common to all predicate.

RESULTS

We have statistically analyzed the dataset by providing the frequency of occurrence of each predicate in the dataset showing the event of the News. We also extracted the frequencies of each predicate as keyword in the dataset. The result of comparison of the predicates to the keyword retrieval is presented in the Figure 2. As it can be observed in the figure that the

ISSN: 2307-3721, e ISSN: 2307-3713

frequency of each keyword is much higher than the SRL predicates, because SRL extracts only if certain word is used as event in the text, while keyword extracts the every occurrence of the word without its context. As in English verbs have different forms, not each form represents the action.

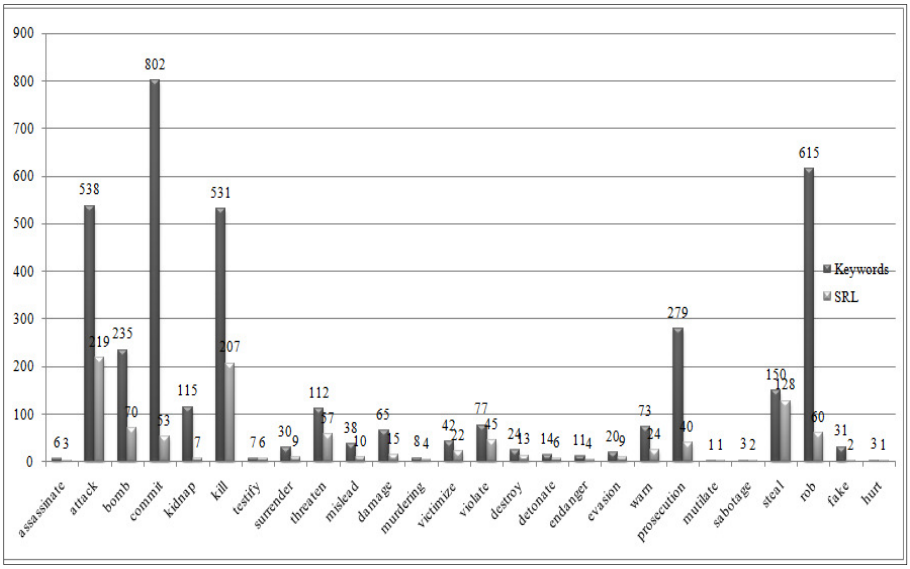

Figure 2. Keywords vs predicates extracted from the dataset

CONCLUSION AND FUTURE DIRECTIONS

This paper presented an efficient method of News analysis. With the proposed method of News analysis, one can efficiently underscore the thrilling aspects of the News reports. Instead of going through the complete News story, with the proposed method the prominent concepts of the News can be dissected swiftly and effectively. The proposed method of News analysis can help News analysts, law enforcement agents and other concerned to glimpse the News promptly and reasonably. The News can be presented using the major concepts/events, entities causing the events or affected by the event as well as location, temporal and instrument information if any is present in the report. We illustrated the major concepts/ events present in the FBI News with the type of entities involved in the events. A statistical analysis of the occurrences of the predicates in the dataset is also presented as well as occurrences of these predicates as keywords. In the current study, we did not manually analyze the context of the keywords in order to further compare them with the SRL extracted predicates. This problem is left as our future work and we aim to manually analyze keywords to further refine the results. In this study we also did not consider live News analysis; therefore it is another point for future direction. Analysis of online News / headlines can be useful for a common News reader who wants to have an instant look on the News.

ACKNOWLEDGEMENTS

Authors would like to thank Anita Miller and James R. (Bob) Johnson ADB Consulting, LLC. University of Texas at Dallas, for providing FBI News dataset. Authors would also like to thank the Masters students Morten Gill Wollsen, Emil Nissen Gaarsmand and Rasmus Frostholm Petersen for parsing FBI dataset using SRL tool.

This work was carried out during PhD study of first author at University of Southern Denmark.

ISSN: 2307-3721, e ISSN: 2307-3713

REFERENCES

[1]. Shen, D., & Lapata, M. (2007). Using semantic roles to improve question answering. *In proceedings of joint conference on empirical methods in natural language processing and computational natural language learning*. Association for *Computational Linguistics*. pp. 12-21

[2]. Avramidis, E., &Koehn, P. (2008). Enriching morphologically poor languages for statistical machine translation. *In proceedings of ACL'08, Association of Computational Linguistics.* pp. 763-770

[3]. Surdeanu et al. (2008). Using Predicate-Argument structures for information extraction. *Proceedings of 41st meeting on Association of Computational Linguistics.* pp. 8-15

[4]. Kulick et al. (2004). Integrated annotation for biomedical information extraction. *HLT-NAACL 2004 Workshop: Biolink 2004*. pp. 61-68

[5]. Carreras, X., & M`arquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic Role Labeling. *Proceedings conference on computational Natural Language Learning (CoNLL-2005), Association for Computational Linguistics*. pp. 152-164

[6]. Gildea, D., & Jurafsky, D. (2008). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245-288.

[7]. Björkelund et al. (2010).*A high-performance syntactic and semantic dependency parser.* In Coling 2010: Demonstration Volume. pp. 33-36.

[8]. Tsai et al. (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, *BMC, 8*(1).

[9]. Haji et al. (2009). *The CoNLL-2009 Shared Task: Syntactic and Semantic* Dependencies in Multiple Languages.

[10]. http://www.fbi.gov/news/stories

[11]. Milos Krstajic, Florian Mansmann, Andreas Stoffel, Martin Atkinson, and Daniel A Keim.

[12]. Processing online news streams for large-scale semantic analysis. *In Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference* on, pages 215–220. IEEE, 2010.

ISSN: 2307-3721, e ISSN: 2307-3713