**A NATURAL LANGUAGE APPLICATION FOR PREDICTION OF HIGHEST RISKS IN A GOOGLE MAP**

# Authors: RUMELA SARAWATI / SUNNY ANAND

***Abstract -* Text mining is a process of extracting information of interest from text, however text mining needs**

**techniques/ concepts from various related domains such as Information Retrieval (IR), Natural Language Processing (NLP), and Information Extraction (IE). In this study, text mining methods and Natural Language Processing Techniques (NLP) are applied to extract causal relations from road accident investigation reports of Dallas Area collected from [1]. The objective of this study is to predict road safety by studying the accident patterns for a cause in a location of a particular area. A careful study of the contributory causes can improve the road safety in an area and in future this concept if integrated with Google map API or a live database can help commuters be more informed about the safety measures on a particular traveling route. The following approach has been employed in this study to extract the causal relationships. Semantic Role Labelling of both ill formed and well -formed sentences, calculating the frequency of the cause of accident for a particular location and then based on the most frequently causing reason the safety prediction is suggested to the user for that location.**

**INTRODUCTION**

There is a growing need for systematic motor vehicle safety efforts which began in the United States in the 1960s. According to a study by [9], average vehicle sales in USA is close to about 18.55 million for October, 2015. The most potential threat to public is the growing number of road accidents. According to [10], a large number of people are victims of road accidents worldwide and even more are disabled by them yearly. However, effective prediction tools for road safety are yet to be developed.

Whenever people travel they are constantly in need of a guide who can tell us the path which they can take to reach a known or unknown destination. In today’s modern world our digital guide is not an unknown entity but Google Map – which guides us to our destination with its utmost precision. However the application which only shows the path to a destination with few other features, has never looked into the aspect of making travelling safe. The inspiration of our research and NLP application is to build a functionality where we make the user of a google map aware of the top reasons of accident and mishaps on the road they are travelling – allowing them to avoid any such mishaps.

The purpose of the project is to develop an application to extract the causes of accidents from accident investigation reports. These causes can be utilized to predict the safety measure to the user accordingly.

**MOTIVATION**

Due to technological advancements in the last century, the automobile industry has gained momentum. Though these technological advancements are beneficial, they have their shortcomings too. Due to increasing number of motor vehicles on the road the number of accidents has also risen. According to [12], the number of victims of road crash accidents worldwide is very huge.

The accidents for USA are recorded by [6] and for Texas by [1]. The board reports how the accident occurred, the causes, location, vehicle, and time. However, they do not provide any recommendations for avoiding similar kind of accident for a particular location in future.

In this project accident investigations are collected from [1] and examined to find the categories to which the accident belongs. The types are classified as Type1 and Type2 based on 2 different feature set (A feature set is a combination of both Syntactic and Lexical features).

Type1 feature set includes a) all accident information collected from various sources using text mining b) events causing the accidents c) the nature of the accident, causal agents d) location of the accident e) vehicle involved in the accident f) time of the accident.

Type 2 feature set includes a) all accident information collected from various sources using text mining b) events causing the accidents c) the nature of the accident, d) causal agents e) location of the accident f) vehicle involved in the accident.

Human intervention is required to extract the causal patterns from the investigation reports as they are unstructured data. The extraction is a complex job as the data is mostly unstructured and ill-formed according to the syntactic and semantic constructs of English Language. Hence these challenges have been attempted to overcome with Natural Language Processing concepts to attempt to create an application close to a real time enhancement package to road safety.

**PROJECT PROBLEM**

The topic of study is a challenging task since the reports have been written in natural language – only English sentences are considered for this study, emphasis has been placed on mining information from unstructured information resources like accident investigation reports.

The problem is formulated as follows:

1. How unstructured data in the form of natural language text-accident reports, be used to derive information about the causes of road accidents specific to a location or place and how can these be used to make users more informed of the severe potential threats at those locations?

The following research questions help solving the research problem under investigation.

1. How the reported accidents are stored and how can these be collected to form the list of all accidents?
2. What categories of accident investigation reports should be considered?
3. How the collected list of accidents be transformed from raw format to computer readable format?
4. How the structured data be then used to derive different features to examine and categorize the accidents based on these derived features?
5. How the features are related semantically and then using those relations to create the underlying inferences of cause and effect?
6. How these causal relationships can then be used to derive on the various road threats and safety actions?

The above questions are answered in this project and a balanced approach is taken to evaluate each options available for each of the above questions.

**PROPOSED SOLUTION**

The proposed solution for the system works for finding and informing the user of the various cautions that need to be followed while travelling on a particular path. As the application is not connected to a live system so the application aims at finding the accidents happened at the locations and then proposing a suitable safety measure

for that location to the user. The idea is to build a knowledge of all accidents which have happened in the past at a location and then using it derive semantically the cause for each accident. These causes are then used to determine which suitable safety guideline to be presented to the user while travelling on that path. The application’s baseline system measures a broad range of feature set to determine all possible and relevant accident reporting to be captured into custom feature set value. The advance system will then determine reducing the feature set and working toward achieving the same level of accuracy as obtained using a full feature set in the baseline system.

**FULL IMPLEMENTATION DETAILS**

The proposed solution is based on a waterfall model of software development.

There are 5 stages in which the application is built.

**STAGE1 – Raw Data collection and Corpus Creation**

Due to unavailability of corpora pertaining to domain of accident reports, the project demanded creation of a new corpora.

1. This corpora was built using text mining technique of Web Crawling. A set of search patterns were created to derive the relevant URL/links which contain the matched information. Examples of such search Pattern /\*Accident Report\*/, /\*Accident in USA\*/, /\*Accident in USA 2014\*/, /\*Accident in Dallas Texas\*/, /\*Accident in Dallas Texas 2014\*/
2. The Result of each search from the patterns mentioned the entire list of Documents (URL’s) were searched to find the Headline of each document using [2].
3. The data was downloaded as an XML file and was then used to extract all the headlines (Link Text from the file) for any relevant information- this created the raw data for this project.

**STAGE 2 – Static Prediction Creation**

The raw data collected from Stage1 is used to determine the relevant predictions.

1. The raw data is split into sentences.
2. The sentences are sentence tokenized using NLTK
3. Sentences are then passed to NLTK’s POS tagger to return the POS tagged corpus.
4. Filter sentence components tagged to different Verb category .We choose a comprehensive list of Verb tag features like “VBG”, “VBN”, “VBD”, “VB”, “VBP”, “VBZ”.
5. Refine the list to find all the unique Verbs and create the dictionary of static predictions for each dictionary entry. The static predictions are defined manually at the present .Example: "Dangerous road ahead! Remain alert!", "Alert! Overpass ahead!"

**STAGE3 – Creation of Structured Data**

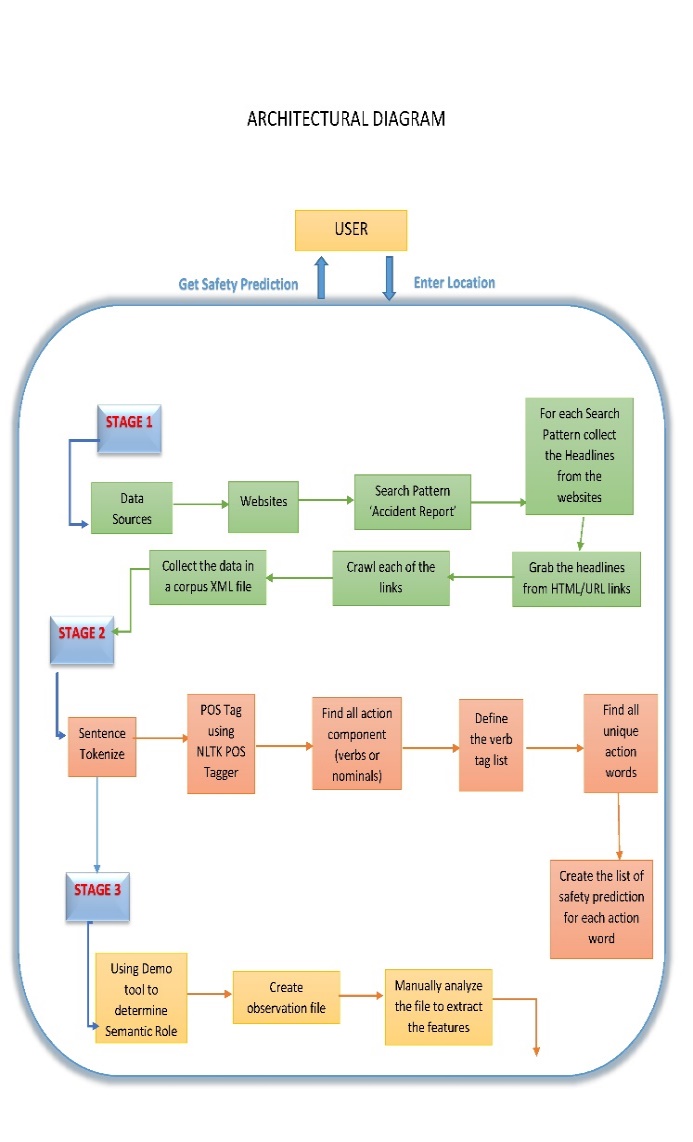
In order to find the causal relation the following methodology has been adopted for our study.

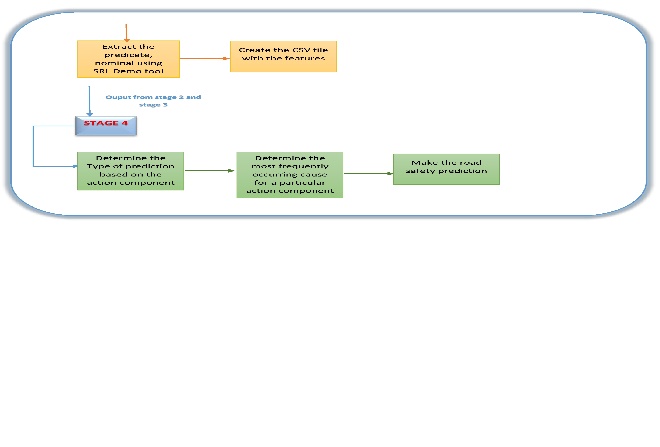
1. To extract the relevant relation the tokenized raw data file is passed through a freely available SRL Tagger [3].
2. There were observations made with respect to the raw data: the collected data was either ill-formed or well-formed sentences in English.
3. The ill formed sentences were discarded and the processing continued for Semantic Role Labelling for well-formed structure using [3].
4. After applying Semantic Role Labelling which is done based on the structure of data in FrameNet in [3], all the passes for a sentence predicate Like SRL(main Verbs),NOM(Nom-Verbs) and Temporal were obtained from [3] and stored for each sentence as an observation in a file.
5. The file above was then analyzed humanly to extract and create a list of features which would be used for creating the relations. Features included location, vehicles, etc.
6. Lemmatizing is done by [3] for the sentences and the corpus is structured into a csv file format
7. Only relevant information like the entire sentence, its verb, the SRL cause, the nominal cause, Location, Vehicle, Temporal and Nominal tags are collected for each sentences different parses for SRL(predicate), NOM and Temporal arguments Semantic Role Labeling [refer 3].
8. All this information is mapped into a manually created CSV file with sentences and feature set values.
9. A determination of the cause and effect of the accident is made using this structured data.
10. After the corpus is made into a readable CSV format, checking is done to see if the cause of the accident can be extracted from the SRL Cause from the verb or the Nominal Cause from the nominal. Once the cause is determined the extracted information is categorized based on various types of available information where Type1 includes all sentence, verb, verb category, cause, location, vehicle, temporal, Type2 sentence, verb, verb category, cause, location.
11. Weights are assigned to the most recurring cause of the accident. The most frequently occurring cause is assigned to be the highest cause at that location.

**STAGE4 – Feature set evaluation and Prediction of Road safety**

1. From the CSV file create the 2 types of categories for determining the cause and location and their count.
2. Type1 feature set includes a) all accident information collected from various sources using text mining b) events causing the accidents c) the nature of the accident, causal agents d) location of the accident e) vehicle involved in the accident f) time of the accident. This creates the list of causes of an accident at a location with their time.
3. Type2 feature set includes a) all accident information collected from various sources using text mining b) events causing the accidents c) the nature of the accident, causal agents d) location of the accident e) vehicle involved in the accident. This creates the list of causes of an accident at a location.
4. From this list in step3 and step2 find the highest occurring cause at a location and find the number of times this cause has led to an accident.
5. For this cause assign the corresponding prediction determined in Stage2.

**ARCHITECTURAL DIAGRAM:**

****

****

**ALGORITHM FOR CAUSE LOCATION AND FREQUENCY OF THE TWO:**

* **From the csv file, read the column headers :**
* **If Cause == SRL cause,**
* **Then, the action component is the predicate**
* **Else If Cause==Nominal cause,**
* **Then, the action component is the nominal**
* **From feature set of each sentence ,** 
  + - **Determine the Types of the categories**
    - **if sentence, verb, cause, location, vehicle, temporal**
      * **cause is Type 1**
      * **create(cause,location)**
    - **Else if sentence, verb, cause, location, temporal**
      * **cause is Type 2**
      * **create(cause,location)**
* **From the (cause,location),for a given location determine the cause by finding the highest frequency cause**
* **Make the prediction based on the cause by its most frequently occurring Type**

**PROGRAMMING TOOLS:**

1. NLTK TOKENIZER, POS TAGGER(NLTK PACKAGE FOR PYTHON)
2. URBANA-CHAMPAIGN SRL TAGGER DEMO TOOL(LIMITED USER)
3. SEO OPTIMIZATION TOOLS[3]
4. PYTHON PROGRAMMING LANGUAGE
5. MICROSOFT CSV FILE for SRL TAGGED DATA FORMATTING

**CHALLENGES**

Several factors have posed to be quite challenging in making the correct predictions based on the cause and effect of an accident in a location. Below are the challenges we faced:

1. <http://www-fars.nhtsa.dot.gov/States/StatesCrashesAndAllVictims.aspx> - These data are collected over the years and we wished to make a prediction upon verification of the data. To do so we had to collect and verify all the accidents reported for a year for a state in US i.e. all the accidents which happened in 2012 in Texas, USA. How can we validate this data? To this we have in mind to use the Google corpus. We want to run over the Google corpus and we want to create a dataset bank for each of the states. Then we plan to classify this to category of road accidents.

We had to discard this approach as Google Corpus is not free.

1. Large part of the data extracted from internet through web crawling is unstructured due to which semantic role labelling was a challenge. The data had to be streamlined to form a corpus of reasonable size in a structured pattern to ensure proper semantic role labelling for which we followed the approach in stage3 of implementation.
2. We also discarded ill-formed sentences.
3. We had challenge in creating the frames for words like “Accident”, we used closely related frames such as catastrophe to resolve it.
4. After semantic role labelling was performed on the corpus, the cause was manually extracted from the verbs or the nominal as opposed to real time data extraction using standard techniques as we needed our own set of features and had defined our own feature tags.

**LIMITATIONS**

The project is limited in scope due to several factors.

1. The idea of creating a real time application could not achieved due to lack of required data set and Google MAP API integration authorization.
2. The size of the corpus is a main setback to predicting the safety measure in real time. We have limited the number of structured sentences after applying Semantic Role Labelling to 30 sentences for training data. Due to this, we have considered only Dallas area for the study.
3. The SRL tagged sentences were not in a format which we can parse for obtaining the relevant information and work on it. So we had to find a new approach to create a new feature set for our application.
4. The improvement upon the baseline system was very difficult due to unavailability of proper data. To show the improvement by reducing the feature set we tried to achieve the same level of accuracy and hence we created 2 types of categories in the project with each category(Type 1 and Type 2) having its own feature set.

**ASSUMPTIONS**

Due to the challenges and limitations the following assumptions have been made:

1. The corpus is Semantic Role Tagged to extract the verbs or nominals to find the cause of the accident.
2. All the sentences are well formed.
3. Accidents captured are motor vehicle accidents.
4. The test data to be used for testing should be in the same format as used in the training data as we had our own training feature set based on which the determination of relations is done.

**POTENTIAL IMPROVEMENTS**

1. If the original corpus and Google authorization is available the same can be used to create a live application.
2. After training the model for the initial data set of the Google Corpus, the application can use supervised machine learning approaches to make predictions and keep adding new heuristics by expanding on the initially created feature set.
3. The static predictions can also be made dynamic based on the new feature set derived in Step II.

**REFERENCES:**

1. <http://accidentdatacenter.com/us/texas/dallas-ft-worth-tx>
2. <http://tools.seochat.com/tools/online-crawl-google-sitemap-generator/>
3. <http://cogcomp.cs.illinois.edu/page/demo_view/srl>
4. <http://naacl2013.naacl.org/Documents/semantic-role-labeling-part-1-naacl-2013-tutorial.pdf>
5. <https://framenet.icsi.berkeley.edu/fndrupal/>
6. <http://www.nhtsa.gov/>
7. <https://en.wikipedia.org/wiki/Epidemiology_of_motor_vehicle_collisions>
8. <http://injuryprevention.bmj.com/content/7/3/176.full>
9. <https://ycharts.com/indicators/auto_sales>
10. http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics
11. <http://www.drivers.com/article/1173/>
12. <https://en.wikipedia.org/wiki/Epidemiology_of_motor_vehicle_collisions>