Road Accident News Information Extraction

Conference Paper · December 2018

CITATIONS
0

READS
11

4 authors, including:

Kritish Pahi
Tribhuvan University
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Road Accident News Information Extraction

Kritish Pahi[1], Aman Shakya[2], Basanta Maharjan[3], Amit Bhandari[4]

Dept. of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering

kritishpahi@gmail.com[1], aman.shakya@ioe.edu.np[2], Basmhzn78@gmail.com[3], amitbhandari@gmail.com[4]

Abstract

This paper describes a system that aggregates road accident information from news feeds extracting information like death, injuries, location, date, day and vehicles involved in the accident. The extracted information describing accidents is maintained in a road accident database system. The approach is based on pre-specified structures called entities, each consisting of number of attributes that are extracted by the Information Extraction System from the news articles. Different subtasks of Information Extraction, Named Entity Recognition, Semantic Role Labelling and Regular expression are employed to extract structured information. The system has a result of 83 % accuracy when tested on one hundred news reports collected from three different sources. The collected information is presented in an online system which can facilitate study of accident patterns.

Keywords: **road accidents, accident news, information extraction, semantic role labelling**

1. Introduction

Road accident is one of the leading causes of death in Nepal. A total of 95,902 crashes, 100,499 injuries and 14,512 deaths were recorded by traffic police over the period of 2001-2013 [1]. However, currently there is a lack of a centralized database of accident data that can be used for analysis. This paper presents a system developed to extract the deaths and injuries in the accidents along with additional entities describing the accident such as location of the accident, vehicles involved in the accident (their Number plates), the day of occurrence (date and day) from the online news. The system processes news from multiple sources in real time to assist people looking for detailed information about the accident. The database maintained would be accessible to all the users (road safety, engineers, police, and statisticians) who want to explore accidents information. Madina [2] has extracted information relevant to user's need from a number of unstructured texts reports about earthquakes occurring throughout the world. Mining crime information from online newspaper articles with Named Entity Recognition (NER) algorithm and Conditional Random Field (CRF) and making the information available to public [3]. Ito [4] LonMaps is information system for a crime and accident mapping system based on news article using a thesaurus and sentences patterns.

2. Methodology

The approach used to implement this information extraction system consists of the following stages:

A. Data Collection & Pre-Processing

Different online news portals were used as the source of road accident news. It was observed that each website had its own unique structure and style of writing. Online RSS Feeds were generated from those sites whose RSS Feed weren't available. Different web crawls were used to crawl the news body from the feed. The news report was then pre-processed by removal of stop-words and stemming and thus reduced sentences were fed into the system to extract information.

B. Name Entity Recognition (NER)

NER was employed for the extraction of location and day of the accident mentioned in the news report.

Training NER:

The custom developed NER model to extract location and day entities was trained with the Groningen Meaning Bank (GMB) corpus. The tags for the NER in the corpus are: geo, org, per, gpe, tim, eve & nat for Geographical Entity, Organization, Person, Geopolitical Entity, Time Indicator, Event and Natural Phenomenon respectively

on the basis of their features. The feature set of the sentence is identified by the word, its lemma, its POS tag and the shape of the word. It is important for the classification to extract the same features for the next-word, next-next-word, previous-word and prev-prev-word to better accuracy. Incremental Learning Classifier was used for learning from batches and implemented using Scikit-Learn Incremental Classifiers.

NER Implementation:

Named Entity(NE) chunker takes Part Of Speech(POS) tagged sentences so that the given input is firstly POS tagged and parsed into the trained NE-chunker parser. Then the previously trained data model is loaded and used for the NER tag extraction.

C. Semantic Role Labelling

Semantic role labelling is the task of automatically finding the semantic roles of each argument of each predicate in a sentence. The sentence-level semantic analysis of text is concerned with the characterization of events, such as determining "who" did "what" to "whom," "where," "when," and "how". [3] Semantics roles are representations that express the role that arguments of a predicate can take in the event like the AGENT. AGENTS tend to be the subject of an active sentence. SRL in action in one of the news body is shown in figure 1.

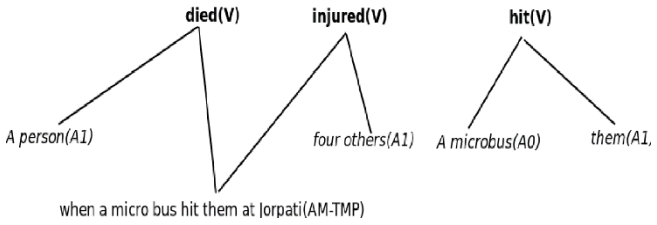

Figure 1. Semantic Parse Tree

"A person died and four others injured when a micro bus hit them at Jorpati." [kathmandupost.ekantipur.com]. Practnlptool is used for Semantic Role Labelling (SRL) which is the implementation of SENNA.

D. Regular Expression Based Information Extraction

Regular expression (regex for short) was used to search pattern in news report in order to extract Vehicles involved in the accident since they follow certain patterns. Along with the number plates of vehicles, the system also identifies the type of vehicles.

3. Entities

The extraction template consists of eight entities which are listed below.

A. Death: Death information is obtained with semantic role labelling (SRL) using death verbs as *die*, *kill*, *crush* and *pass*. The predicate which matches the death verb provides the required argument. Acceptor ('A0') is chosen as the answer whereas in absence of Acceptor, Accepted argument ('A1') gives death information.

B. Death Number: Word to number conversion algorithm is used to convert the death information extracted from SRL into death number.

C. Injury: Injury information is extracted with SRL using injure, sustain, critical, hurt, wound, harm and trauma verbs. Acceptor ('A0') is chosen as the answer whereas in absence of Acceptor.

D. Injury Number: The injury field information is used to derive Injury Number using the word to number conversion algorithm. Injury Number extraction depends on the extracted injury information.

E. Location: Location of the accident is extracted using Named Entity Recognition (NER). The trained model of NER is loaded and used in the extraction of location using "geo" tag.

F. Number Plate: Regular Expression is used to extract the pattern alphabets-digits-alphabets-digits which describes the vehicle number in the news report. Along with the number, the vehicle type is also extracted as word previous to vehicle number is the type of vehicles.

G. Dates: Dates are written in different formats (*yy/mm/dd, dd-mm-yy, July 1, 2017 etc*). Dates are extracted from the beginning of the news report since all the news report follow this pattern.

H. Day: The Day of the accident occurrence is extracted using NER. NER consists of a named tag "tim" which is a time indicator. It indicated the days of the week.
Jaccard similarity is used to detect the duplicate news in the database. Once the entities are extracted from the news, before saving it to the database, the given news text is checked with all the existing record. If any duplicate news exists, it is not stored in the database.

4. Experiments and Results

From three different news sites, a total of one hundred road accidents news were collected and manually annotated and then compared with the result obtained from the system. It was observed that some elements were recognized fully or partially. Hence, not only full matches were considered; credit was also given to partial matches. For partial matches, a coefficient of 0.5 was allocated when half of the elements were extracted. A value of 0.75 coefficient was used in cases where a large part of the element is recognized while a 0.25 coefficient value was used when only a small part of the element is recognized. To valid this, recall, precision and F-measures were calculated for each template extracted and the result is shown in the table 1. The overall accuracy of the system was calculated as below.

$$Accuracy = \frac{COR + 0.75 * PAR + 0.5 * PAR + 0.25 * PAR}{COR + PAR + INC}$$
$$= \frac{40+0.75*53+0.5*7+0}{40+60}$$
$$= 83.25\%$$

where, COR is correctly Extracted elements
PAR is Partially extracted elements
INC is Incorrectly extracted elements
MISS is Missed elements

Table 1. Performance scores for different entities

Entity	Recall (%)	Precision (%)	F-measure
Death	99.5	99.5	99.5
Death No.	94.5	99.47	96.92
Injury	83	92.2	87.35
Injury No.	65.75	91.31	76.45
Date	97.75	97.75	97.75
Vehicle No.	92.6	92.6	92.6
Day	96.66	89.60	93.1
Location	96.25	99.22	97.71

5. Conclusion

This paper presents a methodology for the extraction of key information regarding the road accident from online newspapers from different news sites available. It employs named entity recognition (NER) algorithm, semantic role labelling (SRL) and regular expression for the extraction purpose. The overall accuracy of the system was found to be 83%. Further improvements can be done in the field of visualization and presentation of extracted data. Also, with implementation of deep learning algorithm the current accuracy of extraction can be improved.

6. References

[1] Karkee, R. & Lee, A. H. (2016) *Epidemiology of road traffic injuries in Nepal, 2001-2013: systematic review and secondary data analysis*
[2] Ipalakova, M. (2010) *Information Extraction*
[3] Arulanandam, R., Roy, B. T. & Maryam, A. P. (2014) *Extracting Crime Information from Online Newspaper Articles*
[4] Ito, H. (2014) *LonMaps: An Architecture of a Crime and Accident Mapping System based on News Articles*
[5] Maynard, D., TablanV., Ursu, C., Cunningham, H. & Wilks, Y. (2002) *Named Entity Recognition from Diverse Text Types*
[6] Marquez, L., Carreras, X., Litkowski, K. C. & Stevenso, S. (2008) *Semantic Role Labeling: An Introduction to the Special Issue*
[7] Reschke, K., Jankowiak, M. & Surdeanu, M. (2012) *Event Extraction Using Distant Supervision*
[8] Osoba, O. (2013) *Information Extraction for Road Accident*