Casualty Information Extraction and Analysis from News

Basanta Chaulagain
Pulchowk campus, Institute of Engineering,
Tribhuvan University
071bct510@pcampus.edu.np

Aman Shakya
Pulchowk campus, Institute of Engineering,
Tribhuvan University
aman.shakya@ioe.edu.np

Bhuwan Bhatt
Pulchowk campus, Institute of Engineering,
Tribhuvan University
071bct511@pcampus.edu.np

Dip Kiran Pradhan Newar
Pulchowk campus, Institute of Engineering,
Tribhuvan University
071bct516dip@pcampus.edu.np

Sanjeeb Prasad Panday
Pulchowk campus, Institute of Engineering,
Tribhuvan University
sanjeeb@ioe.edu.np

Rom Kant Pandey
Sanothimi Campus, Tribhuvan University
romkant@gmail.com

ABSTRACT

During unforeseen situations of crisis such as disasters and accidents we usually have to rely on local news reports for the latest updates on casualties. The information in such feeds is in unstructured text format, however, structured data is required for analysis and visualization. This paper presents a system for automatic extraction and visualization of casualty information from news articles. A prototype online system has been implemented and tested with local news feed of road accidents. The system extracts information regarding number of deaths, injuries, date, location, and vehicles involved using techniques like Named Entity Recognition, Semantic Role Labeling and Regular expressions. The entities were manually annotated and compared with the results obtained from the system. Initial results are promising with good accuracy overall. Moreover, the system maintains an online database of casualties and provides information visualization and filtering interfaces for analysis.

Keywords

Casualty, information extraction, news articles, casualty data visualization

INTRODUCTION

This paper proposes a system to extract casualty related information from news reports of any incident. The incident may be any situation of crisis, disaster, accident or attack that involves casualties - deaths and injuries. In particular, news of road accidents has been used for experimentation in this work.

For the past two decades, online news has been one of the most reliable and real-time source of information. This information when transformed into organized knowledge can have great significance. Road accidents are one of the leading cause of casualties 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 (Karkee and Lee 2001-2013). The dataset[1] maintained by the Traffic Police provides statistics of a number of accidents, fatality, serious accidents and

[1] Traffic Police Annual Accident Description, https://traffic.nepalpolice.gov.np/index.php/news/traffic-activities/425-annually- accidental-descriptions

normal accidents every fiscal year.

The existing Accident Database System maintained by the Traffic Police suffers from many deficiencies like under-reporting, inconsistencies, inadequate resources and, time-consuming data collection. Moreover, though government records regarding accidents are the most reliable source, the information is not publicly available in many countries, including Nepal (Bhandari et al. 2017). In situations of crisis, such as road accidents, immediate information is mostly reported by local news agencies only. Hence, local news feeds have been chosen for casualty information extraction in this work. Moreover, the extracted information is collected in a database and made available to the public via an online system with analysis and visualization interfaces.

Relevant entities describing the incident such as death and injury, the location of the incident, vehicle involved in the incident, types of vehicles involved, date and day of occurrence are extracted from news feeds. The extracted casualty information is presented in graphical form for analysis based on location, vehicle types and incident dates. The casualty information extraction system works based on Named Entity Recognition, Semantic Role Labeling and Regular expressions.

In the following 'Related works' section, other approaches in literature are discussed and compared. In the 'Methodology' section, the architecture of the system is presented along with proposed methods to extract casualty information. In the 'Results and Analysis' section, some experimental results are presented along with system snapshots and analysis of the results along with limitations of the system. Finally, the paper is concluded along with a discussion of ongoing and future enhancements.

RELATED WORKS

There are a number of online platforms that apply information extraction techniques in different domains e.g. the Europe Media Monitor (EMM) NewsExplorer and the Thomson Reuters Open Calais. The EMM NewsExplorer[2] automatically gathers news articles from around Europe and separates them into clusters of related articles; identifies names of people, places and organizations for each cluster; identifies name variants by applying name matching techniques to all identified names; and populates a database with the extracted information, learning more about the identified person daily. The Thomson Reuters Open Calais[3] uses Natural Language Processing and machine learning algorithms to extract entities, relations, facts, events and topic code and social tags from submitted texts which are then used for further processing.

Hamborg et al. (2018) developed an open-source, syntax-based system that retrieves an article's main event by extracting phrases that answer the journalistic 5Ws (who, what, when, where and why). Sharma et al. (2013) described 5W1H (who, what, whom, when, where, how) event semantic elements extraction using Semantic Role Labeling (SRL) along with its analysis in their work.

Reschke et al. (2014) have done event extraction using distant supervision approach to extract passenger counts, aircraft types, and other facts concerning airplane crash events. Osoba (2015) developed a system for information extraction of road accidents which extracts information about victims, hospitals and police departments using the designed set of rules that are based on identified patterns. Rules are written in the formalism of the Conceptual Annotations for Facts, Events, Terms, Individual Entities and Relations (CAFETIERE) system. CAFETIERE is an information extraction system designed by a research unit in the School of Computer Science, University of Manchester. (Black et al. 2005).

Arulanandam et al. (2014) describe extracting crime information from online newspaper article by employing Named Entity Recognition (NER) algorithms to identify locations and a Conditional Random Field (CRF) algorithm to classify whether a sentence in the article is a crime location or not. LonMaps (2014) is a work on crime and accident mapping based on news articles, LonMaps, extracts crime/accidents, places, dates and personal names from the news article. LonMaps deals with 4W (what, when, where, and who) of an incident. Places of the incident are displayed on maps and their position is obtained using geocoding. Tanev and Atkinson (2008) presented a real-time news event extraction system for violent and disaster events from online news.

There are plenty of works done on the information extraction through RSS feeds. Han et al. (2009) proposed an efficient algorithm to extract the news article contents from the news pages, that is independent of the page layout. Chy et al. (2014) developed a web crawler to extract useful text from HTML pages of news article contents to construct a full-text-RSS in their work of news classification using Naive Bayes Classifier. Qingcheng and Youmeng (2014) proposed an algorithm to extract content from web pages based on RSS index. They computed the feature of content blocks to obtain the body template with high accuracy.

[2] EMM NewsExplorer, http://emm.newsexplorer.eu/NewsExplorer/readme.html

[3] Thomson Reuters Open Calais, http://new.opencalais.com/

Few works have been done to extract the crisis information from social media posts. Burel and Alani (2018) introduced Crisis Event Extraction Service (CREES), an open-source web API that provides annotations for crisis-related documents, classifies event types (hurricane, flood, etc) and identifies information categories (casualty report, donations and volunteers, etc). The classification was backed by Convolutional Neural Networks (CNNs), which was found to outperform traditional machine learning models. CREES is one of the projects under COMRADES[4], a collective platform for community resilience and social innovation during crisis.

This paper proposes a general system that can be used to extract structured information of casualties from unstructured local news reports in any crisis situation, like disasters and accidents, using Named Entity Recognition, Semantic Role Labeling and regular expressions. Further, locations and involved vehicles are normalized and organized hierarchically. Also, the system facilitates analysis by providing information filtering and visualization interfaces.

METHODOLOGY

In the system, a polling daemon runs daily that polls news from the news sources. The fetched link is checked with the saved list of links that have already been extracted to prevent duplication. The entities from the news article are extracted and stored in a database. Then the extracted information can be filtered and visualized in various forms as needed by the user. The prototype system has been developed in the Django framework in Python and PostgreSQL database was used to store extracted entities.

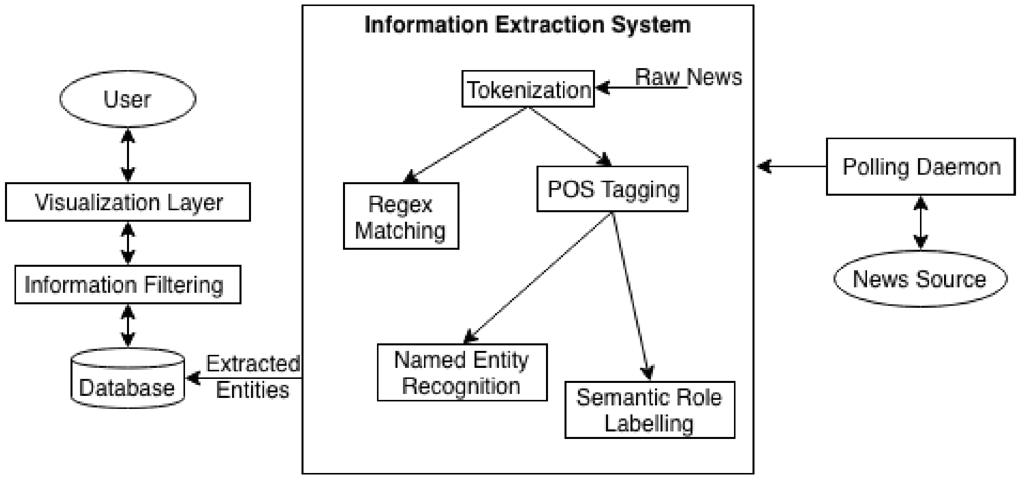


Figure 1. System Architecture.

Data Collection

The online source of news articles used in this work is ekantipur[5], a popular local news site of Nepal. RSS feed of the news articles tagged 'road accident' were fed to the system for information extraction. Since the news website does not support RSS feed, the news is converted to RSS feed using a fetcher named fetchrss[6]. The title and content of each news article from the feed were used.

Usually, news reports follow a common pattern. Every article starts with date of the news which gives the date of the incident. The body usually starts with the casualties involved in the incident specifying the number of people dead and injured along with location where the accident occurred. Vehicles involved are usually mentioned in mid sentences. The latter part usually talks about cause of the accident.

Data Preprocessing

Various methods like tokenization, Parts of Speech (POS) tagging, NER, Lemmatization, Regex matching and SRL are applied on the collected data for the extraction of relevant information.

Sentence tokenization was done using sent_tokenize module provided by Natural Language ToolKit (NLTK).

[4] COMRADES, https://comrades-project.eu/

[5] Ekantipur, http://www.ekantipur.com/

[6] Fetchrss, http://fetchrss.com/rss/5bf76e868a93f84c038b45675bf76e658a93f869028b4567.xml

The module knows what punctuation and characters mark the end of a sentence and the beginning of a new sentence (Loper and Bird, 2002). By looking at the occurrences of those punctuations and sentences the module breaks the paragraph into sentences. Word tokenization was done using word_tokenize module from nltk. It looks for spaces between words to split them.

POS tagging is important to find information regarding vehicle number. It was done using pos_tag module present in the nltk. The module consists of tagger trained in English language which assigns POS tags to the words. To assign tag to a word the POS tagger looks for attributes like if the word starts with a capital letter, has numbers, etc.

NER takes POS tagged sentences as an input to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. It was done using NER tagger from nltk. The tagger is trained on English language and can identify named entities by examining its POS tag and the POS tag of word before it.

Lemmatization refers to the conversion of words form their inflectional form into their base form. Lemmatization was done using lemmatizer from spacy library. To convert a word into its base form the lemmatizer removes suffixes.

News articles usually contain date of news published at the beginning of the first sentence. Due to this POS tagging doesn't work as it should so before extracting day, vehicles number, vehicles involved, etc. the date is extracted and removed from the news article.

Dates are written in different formats (yy/mm/dd, dd-mm-yyy, July 1, 2017 etc). Regex matching the specific format is used to extract date. After extraction the date is removed.

Semantic role labeling, sometimes also called shallow semantic parsing, is a process in natural language processing that assigns labels to words or phrases in a sentence that indicate their semantic role in the sentence, such as that of an agent, goal, or result. It consists of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. The table below represents the meaning of semantic roles for sentences along with the examples.

Table 1. Arguments in an SRL

Thematic Role	Definition	Example
AGENT	The volitional causer (subject of a sentence) of an event	*the waiter* spilled the soup.
FORCE	The entity that instigates an action, but not consciously or voluntarily.	*the wind* blows debris from the mall into our yards
THEME	The participant most directly affected by an event	only after Benjamin franklin broke *the ice.*
RESULT	The end product of an event	The city built a *regulation-size baseball diamond.*
INSTRUMENT	An instrument used in an event	he poached catfish, stunning them *with a shocking device*
SOURCE	The origin of the object of a transfer event	I flew in *from Boston.*
GOAL	The destination of an object of a transfer event	I drove to *Portland.*

An example of SRL is given below.

Input: “A person died, and four others injured when a micro bus hit them at Jorpati.”

SRL output: [‘A1’: ‘A person’, ‘V’: ‘died’, ‘A0’: ‘A microbus’, ‘V’: ‘hit’, ‘A1’: ‘them’, ‘AM-TMP’: ‘when a micro bus when a micro bus hit them at Jorpati’]

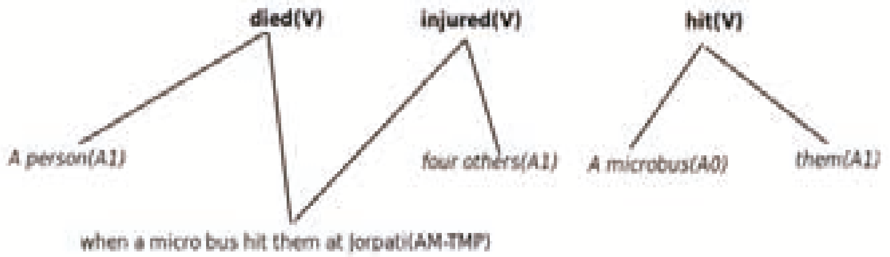


Figure 2. Semantic Parse Tree

Here 'V' refers to the verb, 'A0' refers to the agent and 'A1' refers to patient associated with the verb. AM-TMP is a temporal argument which shows when an action took place. The tag 'V', 'A0' and 'A1' appear in SRL output but they are not semantic roles.

Entities Extraction

The extraction template consists of eight entities which are described below. Entities refer to various fields that are extracted from the news article such as date, vehicle number, number of deaths/injuries, etc.

Death

The Death Field gives information regarding the people who died in the incident. To obtain such information, SRL is used with the death verbs. Death verbs are verbs like die, kill, crush and pass that are present in sentences containing information about death of people. 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. Death information is a phrase containing information about the number of deaths. In figure 2, A1 is associated with death verb (die) and A0 is not related with death verb hence A1 is chosen as the phrase related with death and is used for death number extraction.

Word to number conversion algorithm is used to convert the death information extracted from SRL, in previous step, into death number. The death number extraction depends on the extracted death information from the news. During conversion articles ('A', 'An' and 'The') are converted to number one.

For example: "A person died and four others ...". Here the word "A person" is converted to one and hence number of deaths equals one.

Injury

Injury Field gives information regarding the injured people in the incident. Mapping of injury template is carried out with SRL using injury verbs. Injury verbs are verbs like injure, sustain, critical, hurt, wound, harm and trauma that are present in sentences containing information about the injured people. The predicate which matches the verb provides the required argument. Acceptor ('A0') is chosen as the answer whereas in absence of Acceptor, Accepted argument ('A1') gives the injury field. In figure 2, A1 is associated with injury verb hence A1 is chosen as the phrase related with injury.

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

Location

The location of the incident is extracted using NER. The word that is tagged as 'GPE' by NER is the extracted location. The locations of Nepal are further organized in a hierarchy. A location tree is constructed with the root node as 'Nepal', and 75 districts as its child nodes. Further, under 'Kathmandu' district node, 150+ locations are defined.

The extracted location is tallied with the location in location tree and the location that matches with the highest degree of similarity is stored. Similarity is obtained by calculating Levenshtein edit-distance between the locations. Similarity check ensures that same location with multiple spellings like 'Koteshowr', 'Koteshwor', 'Koteshor' are stored only once as defined in the location tree. In our tree, we've defined 'Koteshwor' and whenever one of the above three locations is extracted, it checks for similarity and as it passes the threshold of

70%, it is stored as ‘Koteshwor’ in the database. The 70% threshold was chosen by experimentation.

Vehicle Number

After POS tagging, Regular Expression is used to extract the pattern *<CD>*<CD>, the asterisk represents any character/group of character and <CD> represents a cardinal number which describes the vehicle number in the news report. Vehicle number in Nepal follow a specific pattern <word><number><word><number> and can be captured using the pattern *<CD>*<CD>.

Example: “The bus (Ba 2 Kha 4683), en route to Bardia from Kathmandu, overturned near the Hugdi bridge, said police”. Here the bus number, Ba 2 Kha 4683, can be captured using pattern mentioned above.

Vehicle specifies the type of vehicle (bike, car, bus) involved in the incident. For vehicle extraction a gazetteer named ‘vehicles' was constructed. The gazetteer contains only lemmatized forms of vehicles. For example, it contains ‘bus’ but not ‘buses’. To find the vehicles, a complete news story is searched to find words whose lemmatized form match the word present in the gazetteer(vehicles) and then the vehicle is categorized into one of three categories two-wheeler, three-wheeler or four-wheeler.

Day

Almost all the news article specifies the day in which the incident occurred in the first few sentences of the news story. Since all seven days of the weekend with the suffix–day we use regex to find the day of the incident.

RESULTS AND ANALYSIS

Results

The extracted information is collected in database as structured data for analysis and visualization. A prototype system has been deployed in a webserver[7]. The site displays the extracted information from the news articles and visualizes it in various forms like graphs, charts and maps.

The main page lists the latest news articles related to road incidents and the extracted items beside the selected news article. Visualization can be done based on the number of incidents and casualties(deaths and injuries). There are filters for location, vehicle types and time range for flexible analysis of incidents.

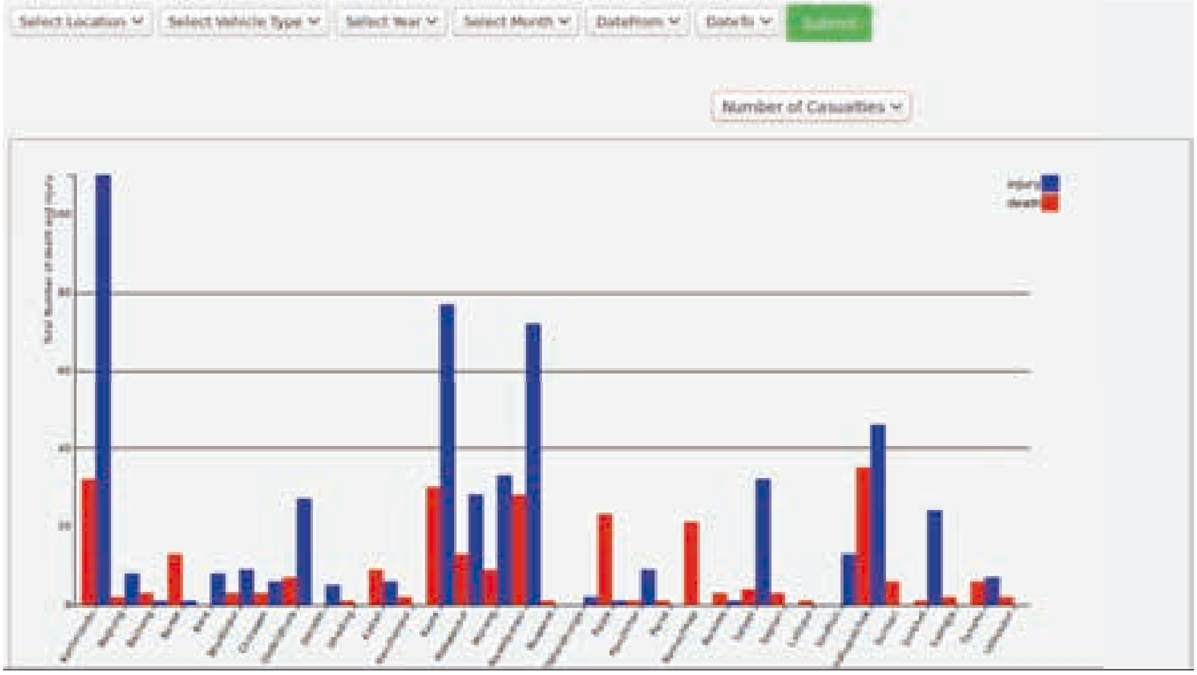


Figure 3. Graph showing district wise casualties.

[7] http://103.5.150.17

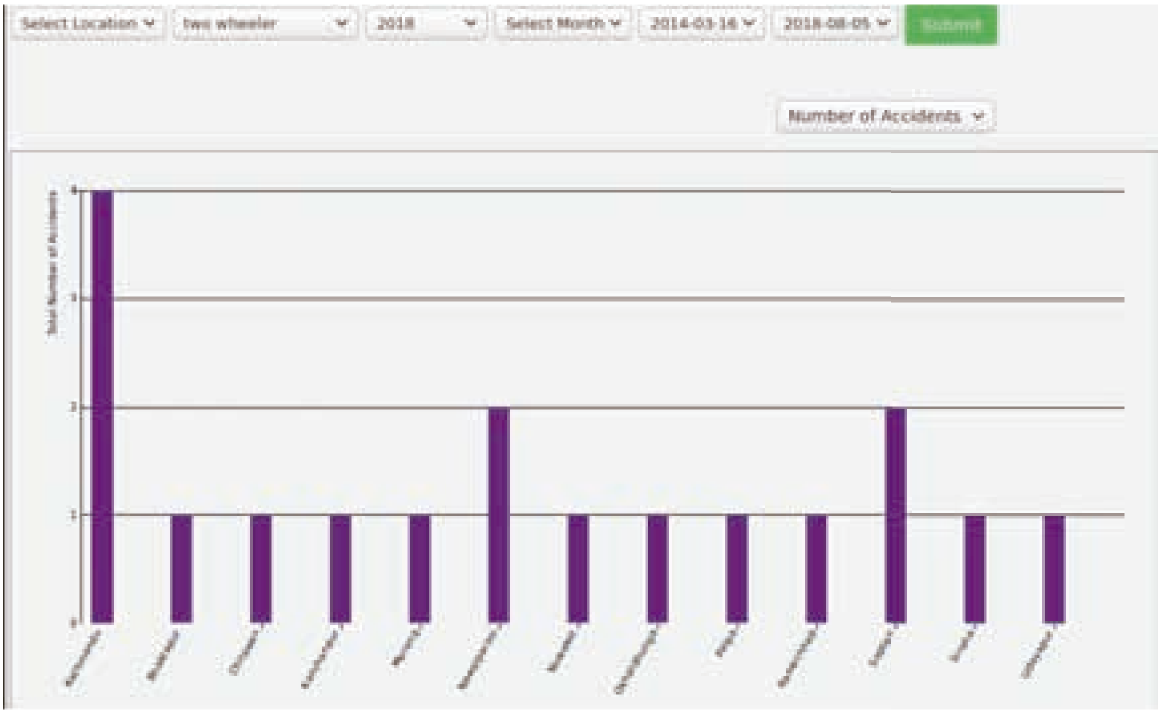
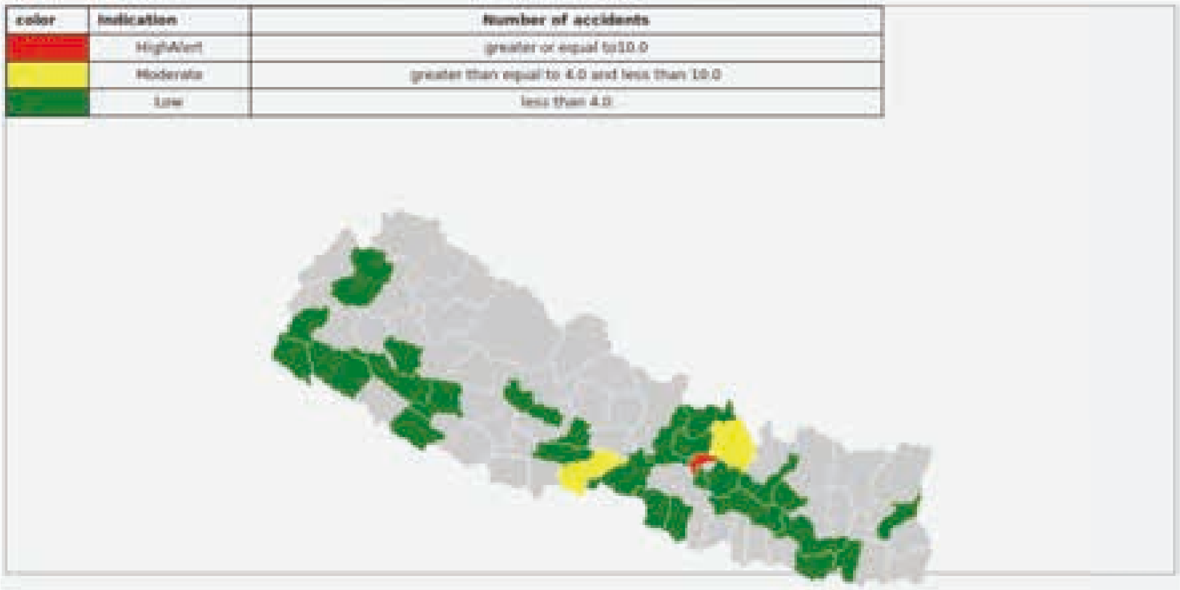


Figure 4. Graph showing district wise two-wheeler incidents in 2018.

Visualization is also possible in the map of Nepal (divided into 75 districts) and map of Kathmandu. In the map of Nepal, the number of incidents is displayed for different districts with three colors: red, yellow and green representing the frequency of incidents, red being the high alert districts. Data from the server is passed in the json format which is processed using D3.js. The threshold value for the alert categories is dynamic and is adjusted according to the total number of accidents.



color	Indication	Number of accidents
🟥	HighAlert	greater or equal to10.0
🟨	Moderate	greater than equal to 4.0 and less than 10.0
🟩	Low	less than 4.0

Figure 5. Map of Nepal showing district wise incidents.

Google Map's API is used to display the specific locations of incidents in the map of Kathmandu. Geocoder was used to obtain the latitude and longitude of a place that was extracted. These values are used to create heatmap using Google Maps API.

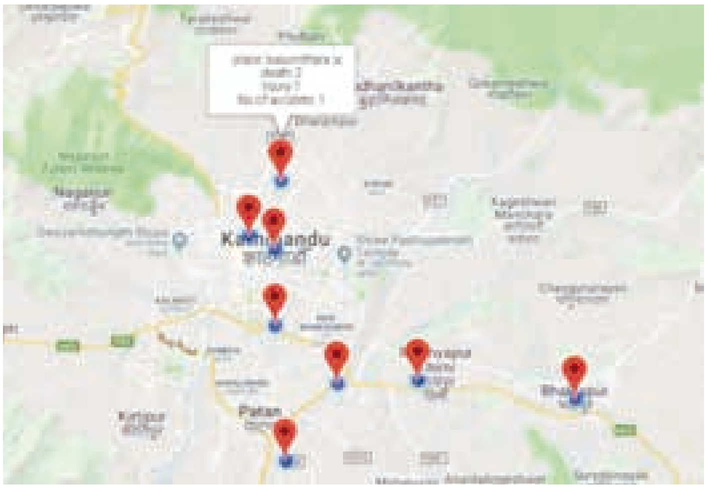


Figure 6. Map of Kathmandu pointing the places of incidents.

Evaluation

The metrics used to evaluate our extraction system is accuracy. Ten entities were used in the analysis, namely death, injury, death number, injury number, location, vehicles involved, vehicle numbers, vehicle types, date and day. The individual accuracy of each extracted entities, the accuracy of each news articles and the overall accuracy of the system is measured. These measures help to test the validity and performance of the designed system.

For an unbiased evaluation of the designed system, 50 news articles were randomly selected. The entities were manually annotated by a group of 10 engineering students and compared with the results obtained from the system to ensure the rules extract the right information. For each entity, the number of correct extractions over 50 news articles was obtained and used in the below equation to calculate the accuracy of individual entities.

$$Accuracy\ of\ entity = \frac{Number\ of\ entities\ with\ correct\ extraction}{Total\ number\ of\ entries}$$

To calculate the accuracy of individual news entries, the two sets of entities, manually annotated and system extracted were compared. The number of correctly extracted entities was obtained and used in below equation to calculate the accuracy of all 50 news entries.

$$Accuracy\ of\ article = \frac{Number\ of\ correctly\ extracted\ entities}{Total\ number\ of\ entities\ in\ an\ article}$$

Finally, the overall accuracy was calculated using the equation below.

$$Overall\ accuracy = \frac{Sum\ of\ accuracies\ of\ all\ news\ articles}{Total\ number\ of\ news\ article}$$

The accuracy of each entity over 50 news articles is as shown in the table below.

Table 2. Accuracy of the Extracted Entities

Entity	Accuracy
Death	94%
Death number	96%
Injury	84%
Injury number	94%
Date	100%
Day	96%
Location	80%
Vehicles involved	88%
Vehicle numbers	96%
Vehicle types	88%

The overall accuracy of the system was found to be 90.8% .

Analysis and Limitations

The system finds the sentences containing pre-defined death and injury verbs and assigns the subject of such sentences as death and injury. The subject of such sentences may not always be the actual deaths and injuries, leading to inaccurate extraction. Similarly, death and injury numbers extracted by the system is not always accurate as it may take other numerals in the sentence as death and injury numbers.

Extraction of date and vehicle numbers is quite accurate as they follow specific patterns while expressing. Extraction of location is least accurate (80%) because there are many locations specified in a news article like the place of incident, the destination of the vehicle, the home address of the victim, the location of the hospital, etc. and system may extract the incorrect location. Also, 75 districts and 150+ places inside Kathmandu are only defined in the location tree. If the extracted location is not in the location tree, extraction is incorrect.

Extraction of involved vehicles is not always accurate as the extracted vehicles should match from the list of pre-defined vehicle gazetteer. Some news articles do not specify the vehicle or have unconventional vehicle names like 'Toyota vehicle', due to which extraction fails. Each vehicle in vehicle gazetteer is categorized as two-wheeler, three-wheeler or four-wheeler, so vehicle type has the same accuracy as vehicles involved.

CONCLUSION AND FUTURE WORK

This paper presented a methodology for the extraction of casualty information from online news articles and demonstrated an online system for the collection, analysis and visualization of the extracted information. Named entity recognition (NER), semantic role labelling (SRL) and regular expression have been used for the extraction purpose. Vehicle and Location extraction have been standardized by using simple vehicle and location hierarchies. The overall accuracy of the system calculated over 50 news articles was found to be 90.8%. Although the system has been implemented for road accidents in Nepal, it can be used in crisis for information extraction from any news involving casualties.

There are still rooms for improvements in terms of accuracy and completeness of the system. Some ongoing and future works are listed below.

- Automatic classification of casualty related news from general RSS news feeds.
- Aggregation of news from multiple sources along with identification and consolidation of duplicate news.
- Updates on the same incident in successive news reports (for example, increase in number of casualties few days after an accident).
- Extraction of the exact cause of the incident for extensive analysis.
- Validation and correction of extracted information to maintain an accurate and up-to-date database.
- Exploration of sentence structures and verbs intention to make more sense from the news.

ACKNOWLEDGEMENT

This work has been supported by the University Grants Commission, Nepal under a Collaborative Research Grant (UGC Award No. CRG-74/75-Engg-01) for the research project "Establishment of a Disaster Telecommunications Research and Educational Facility Advancing a Scientifically Sound Disaster Telecommunication Infrastructure and Processes in Nepal".

REFERENCES

Arulanandam, R., Savarimuthu, B. T. R. and Purvis, M. A. (2014). Extracting Crime Information from Online Newspaper Articles.

Bhandari, A., Maharjan, B. and Pahi, K. (2017). Road Incident News Information Extraction

Black, W.J., McNaught, J., Vasilakopoulos, A., Zervanou, K., Theodoulidis, B., and Rinaldi, F. (2005). "CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations", *Parmenides Technical Report* TR-U4.3.1.

Burel, G. and Alani, H. (2018). Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media. *15th International Conference on Information Systems for Crisis Response and Management*, Rochester, NY, USA.

Chy, A. N., Seddiqui, M. H. and Das, S. (2014). Bangla news classification using naive Bayes classifier. *16th Int'l Conf. Computer and Information Technology.*

Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T. and Gipp, B. (2018). Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions, *Transforming Digital Worlds. iConference 2018. Lecture Notes in Computer Science*, vol 10766. Springer, Cham.

Han, H., Noro, T. and Tokuda, T. (2009). An Automatic Web News Article Contents Extraction System Based on RSS Feeds.

Karkee, R. and Lee, A. H. (2001-2013). Epidemiology of road traffic injuries in Nepal.

LonMaps, H. I. (2014). An Architecture of a Crime and Accident Mapping System based on News Articles.

Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit.

Osoba, O. (2015). Information Extraction for Road Accident Data.

Qingcheng, L. and Youmeng, L. (2014). Extracting Content from Web Pages Based on RSS. *16th Int'l Conf. Computer and Information Technology.*

Reschke, K., Jankowiak, M. and Surdeanu, M. (2014). Event Extraction Using Distant Supervision.

Sharma, S., Kumar, R., Bhadana, P. and Gupta, S. (2013). News Event Extraction Using 5W1H Approach & Its Analysis. *International Journal of Scientific & Engineering Research*, Volume 4, Issue 5, May-2013.

Tanev, H., Piskorski, J. and Atkinson, M. (2008) Real-Time News Event Extraction for Global Crisis Monitoring. *Natural Language and Information Systems. NLDB 2008. Lecture Notes in Computer Science*, vol 5039. Springer, Berlin, Heidelberg.