Extracting Crime Information from Online Newspaper Articles

Rexy Arulanandam **Bastin Tony Roy Savarimuthu** **Maryam A. Purvis**

Department of Information Science
University of Otago,
PO Box 56, Dunedin, New Zealand 9016,
Email: (rexy.arulanandam,tony.savarimuthu,maryam.purvis)@otago.ac.nz

Abstract

Information extraction is the task of extracting relevant information from unstructured data. This paper aims to 'mine' (or extract) crime information from online newspaper articles and make this information available to the public. Baring few, many countries that possess this information do not make them available to their citizens. So, this paper focuses on automatic extraction of public yet 'hidden' information available in newspaper articles and make it available to the general public. In order to demonstrate the feasibility of such an approach, this paper focuses on one type of crime, the theft crime. This work demonstrates how theft-related information can be extracted from newspaper articles from three different countries. The system employs Named Entity Recognition (NER) algorithms to identify locations in sentences. However, not all the locations reported in the article are crime locations. So, it employs Conditional Random Field (CRF), a machine learning approach to classify whether a sentence in an article is a crime location sentence or not. This work compares the performance of four different NERs in the context of identifying locations and their subsequent impact in classifying a sentence as a 'crime location' sentence. It investigates whether a CRF-based classifier model that is trained to identify crime locations from a set of articles can be used to identify articles from another newspaper in the same country (New Zealand). Also, it compares the accuracy of identifying crime location sentences using the developed model in newspapers from two other countries (Australia and India).

Keywords: crime mining, information extraction from newspapers, machine learning

1 Introduction

With the advent of the Internet, huge volumes of data (also called 'big data') are available online. Electronic newspapers are increasingly being read by users from anywhere, anytime. In New Zealand alone there are about 20 daily newspapers, and many of them make an electronic version available online. Newspapers are a source of (mostly) authentic and timely information. There is a large amount of information available in newspaper articles. For example, newspaper articles contain information about crimes, accidents, politics, cultural events and sports events.

Even though valuable information is available in human-readable form in online newspapers and electronic archives, software systems that can extract relevant information and present these information are scarce and this has been of significant interest to researchers in the field of Information Extraction (Cowie & Lehnert 1996). Even though search engines can be used to query specific information (e.g. cultural events in Auckland), these query results do not provide a historical perspective (i.e. if there are 100 articles on cultural events, the user may have to read all of these in order to gain some insights such as the increase in number of operas in a city). Although, one could manually read through the results and extract valuable information, this process is tedious and error prone. So, this work aims to 'mine' information available in online newspaper articles.

In this work, crime information extraction is chosen as the domain for investigation because crime is one of the key variables for people to decide whether to move to a new country (or relocate to a new city) and places to avoid when one travels. For example, a new immigrant may want to compare different cities based on crime rates or compare different neighbourhoods of a particular city to choose a safer one. A traveler may want to know which parts of a particular city to avoid. Currently, this information is not readily available for users, but these can be obtained from newspaper articles of a particular region, as they tend to report the important crimes. To demonstrate the viability of the approach for automatic extraction of crime information, the domain of *theft* has been chosen in this work. Theft information can be extracted at different levels (street level, suburb level and city level) and this information can be visualized on top of maps (e.g. Google Maps). The system that automatically extracts and presents such information has the potential to be used by the residents of various cities to undertake proactive 'neighbourhood crime watch' initiatives to reduce crime. It can also be used by potential immigrants and the visitors of the city to make informed decisions about where to live/stay and also take appropriate precautions when visiting certain areas. Additionally, city councils may use the system to identify crime hot-spots and then employ appropriate monitoring and controlling mechanisms.

In certain countries, crime information is available to the public on top of Google maps. For example UK government has crime map (UK Police 2013) available to the public. However many countries make only the coarse-grained crime information available to the public. For example, crime information in New Zealand (NZ Government 2013) contains coarse-grained information (e.g. total number of thefts in a

district or a province) which is not very useful to an average citizen. Individuals require fine-grained information (e.g. thefts in a particular suburb). Thus, the aim of this research is to extract crime information available in newspaper articles and make the hidden information publicly available. To achieve this goal, this paper discusses a methodology that consists of seven steps. It uses Named Entity Recognition (NER) to identify locations and then employs Conditional Random Field (CRF) to classify whether a sentence containing a location is a crime location sentence[1].

Articles related to theft from newspaper articles over a period of time from four different English newspapers from three different countries were considered for this study (Otago Daily Times from New Zealand, New Zealand Herald from New Zealand,Sydney Morning Herald from Australia and The Hindu from India). The main contributions of this study are three fold. First, it presents a methodology for extracting crime information from newspaper articles. Second, it compares four NER techniques on their ability to identify location information. Third, it evaluates how well the classifier model created for labelling sentences from one English newspaper in New Zealand can be used for identifying theft location information from newspaper articles from three other newspapers, one from New Zealand and one each from Australia and India.

This paper is organised as follows. Section 2 provides an overview of the related work and techniques employed in the area of crime information extraction. Section 3 presents the methodology employed to identify sentences with locations and also to classify whether a sentence is a crime location sentence. Section 4 discusses the various experiments that were conducted using four different newspaper articles from three countries and the results. Section 5 discusses the merits and limitations of the work reported in the paper and also points towards future work. The conclusions are provided in Section 6.

2 Background and Related Work

This section provides an overview of the related work in the domain of crime extraction. It also provides a brief background on the two techniques, Named Entity Recognition (NER) and Conditional Random Fields (CRF) used in this work.

Crime monitoring and prevention is a domain of interest to all countries across the globe in order to make the world a safe place to live. Use of ICT technologies for this purpose has been around since the advent of computers. Currently, with massive amounts of data being available on an individual's activities (e.g. Twitter and Facebook) and the wide spread availability of news articles (e.g. through freely available online newspapers and YouTube), researchers have become interested in combining these information to monitor and prevent crimes. While an individual's activities can be private, most newspaper articles are public. In this work we consider such publicly available information. Also the focus of this work is to identify crime locations (in particular theft crime) and make this available to the public. This information can potentially be used in crime prevention.

Researchers working in the area of crime information extraction have used several techniques. In particular, researchers have used techniques such as crowd sourcing, data mining and machine learning for this purpose. The Wiki Crimes project (Wiki Crimes 2013, Furtado et al. 2010). harnesses the power of the crowd, where individuals report crime details online and other users can use this information to make decisions. However, a limitation of this approach is the difficulty of verifying the authenticity of the posted crimes.

Researchers have explored techniques for retrieving relevant information from unstructured documents. The process of extracting information from unstructured documents is difficult because it is written in natural language and the structure of the document is not known ahead of time (when compared to structured files such as databases). However, there has been a lot of work on identifying entities (e.g. person, place, organization) from unstructured documents in the field of natural language processing. Often called as Named Entity Recognition (NER), this technique has been shown employed in many domains (for an overview see (Nadeau & Sekine 2007)). For example, the Coplink project (Chen et al. 2004) of researchers at the University of Arizona aims at identifying information about criminals from police reports. It uses an Entity Extractor system that is based on AI techniques, for detecting identities of criminals automatically and also for analyzing criminal networks using clustering and block modeling.

There are other works that extract relationships between variables available in the form of structured information (e.g. identifying relationship between column variables of a database table) using data mining techniques such as cluster analysis. For example, the work of De Bruin et al. (2006) uses such an approach for analyzing criminal careers. Based on the analysis, they have identified four important factors (crime nature, frequency, duration and severity). By using these factors, they created criminals' profiles and compared each criminal with all the other criminals using a new distance measure and also clustered similar criminals. They obtained data from the Dutch National Criminal Record Database for their study. Chandra et al. (2007) have employed clustering to identify crime hot-spots based on Indian crime records. These works are mainly based on structured data. In contrast to these works, the work reported in this paper aims to extract information from unstructured newspaper articles.

Researchers have used Conditional Random Fields (CRF), a statistical modelling technique for machine learning. Conditional Random Fields are used to build probabilistic models that can be used to label data. It is a discriminative probabilistic model and it learns weights between features from the training dataset and outputs a *model* that can be used to assign labels for test data (for an overview see (Lafferty et al. 2001)). They have been shown to offer several advantages over Hidden Markov Models (HMMs). Also, they avoid the labelling bias problem suffered by the Maximum Entropy Markov Model (MEMM). CRFs have been used in many domains. For example, Angrosh et al. (2010) have used CRFs for classifying sentences in a research article into different categories (e.g. background sentence, related work sentence, and shortcoming sentence). Peng & McCallum (2006) have also used CRFs to extract information from research papers. The work reported in this paper employs NER algorithms to identify locations and a CRF algorithm to train a model based on a set of features defined by the authors which is subsequently used to assign labels to sentences. Ku et al. (2008) aim to extract crime information from a variety of sources such as police reports, victim state-

[1] A location mentioned in an article does not mean it is a crime location. The objective here is to identify whether the location mentioned in a sentence is a crime location or not.

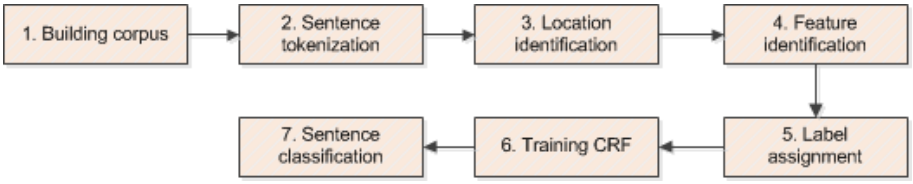

Figure 1: Steps used in the methodology employed

ments and newspaper articles. Among other types of crime information, they identify locations. Their focus is to identify just the locations and not whether the location is indeed a crime location (i.e. there could be other types of locations such as victim's or offender's hometown which may not be the crime location).

Our work is inspired by the approach used by Angrosh et al. (2010). The domains of interest are different in both works, hence, the features identified and labels used are distinct. Our work employs NERs for location identification, which was not required in the cited work because their domain of interest was in the area of labelling sentences in research articles based on their purpose (e.g. background sentence and shortcoming sentence) and does not involve locations details. Also, to the best of our knowledge our contribution is unique to the domain of crime information extraction.

3 Methodology

The objective of this work is to identify the theft location from a corpus and categorize each sentence in an article into Crime Location Sentence (CLS) and Not a Crime Location Sentence (NO-CLS).

This section describes the methodology used for classifying sentences in newspaper articles into CLS and NO-CLS sentences. Figure 1 shows the steps used in the methodology followed in this work.

1. ***Corpus building -*** The first step is to build a corpus of relevant newspaper articles for our study. We used Mozenda Web Screen Scrapper tool (Mozenda 2013) for this purpose. Mozenda is a web scrapping tool to extract the specific information from websites. If we train the tool by pointing and clicking at details that need to be extracted, the tool can extract the same set of information automatically from a given pool of documents. It also saves the extracted information in different formats for latter use. We built a corpus of theft-related articles.

2. ***Sentence tokenization -*** Upon extracting relevant newspaper articles, the individual sentences need to be extracted (i.e. an article needs to be splitted into individual sentences). We used a PunktTokenizer (Kiss & Strunk 2006) from the NLTK toolkit (Bird et al. 2009) for this purpose. The tokenizer divides a given article into a list of sentences.

3. ***Location identification -*** Upon extracting individual sentences, the locations in each of the sentences have to be identified. Locations are identified using Named Entity Recognition algorithms (Nadeau & Sekine 2007). These algorithms are discussed in Section 4.1.

4. ***Feature identification -*** The fourth step is to define a set of features that can then be used to assign labels to sentences. The sentences will be labelled as crime location sentence(CLS) and not a crime location sentence (NO-CLS). The features include *sentHasCrimeTerm* which means that a sentence has a crime term, *sentHasCityLoc* which means that a sentence has city location etc. A list of terms used in features and their descriptions are provided in Table 2. Also, Table 2 provides examples of sample terms (or phrases) that are used to represent a feature and the number of such terms identified from a corpus of 70 articles from Otago Daily Times (discussed in Section 4). For example, the first row shows that the theft related crime terms include theft, burglary, and stolen. There were a total of 336 instances that were identified in the corpus. There were 55 unique terms defined for this purpose (in brackets)[2].

5. ***Label assignment -*** Once features have been identified, the labels are manually assigned for each sentence in an article. The assigned labels are a) CLS - Crime Location Sentence and b) NO-CLS - Not a Crime Location Sentence. Figure 2 shows the sample Crime Location Sentence (CLS) and Not a Crime Location Sentence (NO-CLS). The first sentence has details about the theft (i.e. a car was stolen), the address of the theft crime (i.e. Norfolk St) and the city (i.e. Dunedin). The second sentence only has a street location. A human reading these sentences can classify the first sentence as a sentence that contains a location that is the crime location. However, the second sentence does not provide a clue whether the sentence is a crime location. In order for a system to classify a sentence into a crime location sentence, we need to assign features. It can be observed from Figure 3, three features have been defined for sentence one and one feature has been defined for sentence two. The feature extracting is automated by a python program using regular expressions which assigns features for each sentences. Once the features are assigned for the training data, each sentences has to be manually labelled[3]. The first sentence with the three features will be labeled as CLS and the second sentence will be labelled as NO-CLS. Some sample features of sentences and their labels are shown in Table 3.

6. ***Training CRF -*** From a dataset of articles that have been annotated with features and labels, certain percentage (e.g. 70%) is chosen as the training data. The CRF algorithm learns weights between features from training dataset and creates a *model*. This model, when given a new set of data (e.g. a new article with features),

[2]The values in brackets are absent in many because in those cases unique terms are not defined manually. These are location information (city, suburb etc.). They are obtained through regular expressions on sentences that are annotated with results from NERs.

[3]In supervised learning, a dataset is divided into two parts, training data and test data).

No.	Features	Description
1.	sentHasCrimeTerm	Sentence has crime term
2.	prevsentHasCrimeTerm	Previous sentence has a crime term
3.	sentHasRegionLoc	Sentence has region location
4.	sentHasCityLoc	Sentence has city location
5.	sentHasSuburbLoc	Sentence has suburb location
6.	sentHasStreetLoc	Sentence has street location
7.	sentHasPersonLoc	Sentence has person location
8.	sentHasPoliceLoc	Sentence has police location
9.	sentHasCourtLoc	Sentence has court location
10.	sentHasLocation	Sentence has other country names

Table 1: Features defined and their descriptions

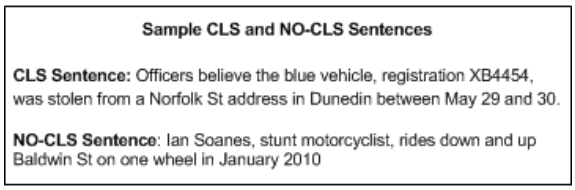

Figure 2: Sample CLS and NO-CLS sentences

automatically assigns labels to each of the sentences in the article. The model produced is used in the next step.

7. ***Sentence classification -*** Once the model is created, it is used to label the sentences (i.e. automatic label assignment as opposed to manual label assignment in Step 5) in the remaining articles (i.e. the test data). The labels obtained through the model are then compared with the labels assigned by the humans. We compute the precision, recall, f-score and accuracy for the results obtained. The formulae to compute these four are given below where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives respectively.

$$Precision(P) = TP/(TP + FP) \quad (1)$$

$$Recall(R) = TP/(TP + FN)) \quad (2)$$

$$F - score = 2PR/(P + R) \quad (3)$$

$$Accuracy(A) = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

The steps listed in the methodology are at a high-level of abstraction. We describe them in more detail in the context of the experiments conducted in the next section.

4 Experiments and Results

We conducted four experiments to demonstrate the efficiency of the system designed to identify crime location sentences. In the first experiment, we studied

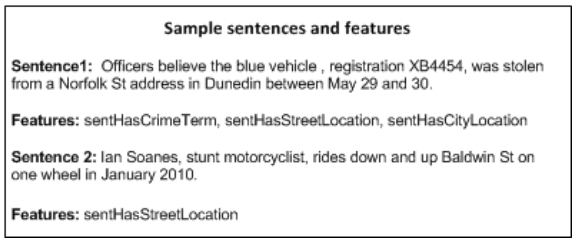

Figure 3: Sample sentences and features

Features of a sentence	Label
sentHasCrimeTerm, sentHasPoliceLoc	CLS
sentHasCrimeTerm, sentHasSuburbLoc, sentHasCityLoc	CLS
sentHasCrimeTerm, sentHasStreetLoc, sentHasCityLoc	CLS
sentHasCrimeTerm, sentHasCityLoc	CLS
sentHasStreetLoc	NO-CLS
sentHasCrimeTerm	NO-CLS
sentHasCityLoc, sentHasSuburbLoc	NO-CLS
sentHasCrimeTerm, sentHasRegionLoc, sentHasSuburbLoc	CLS

Table 3: Sample features and labels

the impact of four types of NER algorithms which identify locations on the classification obtained in a regional newspaper in New Zealand. Second, based on the locations identified by the best model, we evaluated the performance of our system on the accuracy of identifying crime location sentences from the articles in a regional newspaper (Otago Daily Times[4]). Third, using the model created from Otago Daily Times articles, we labelled articles from another newspaper in New Zealand (New Zealand Herald[5]) and investigate the accuracy of the developed model. Fourth, using the same model, we classified (i.e. labeled) sentences from newspaper articles from two countries, Sydney Morning Herald[6] from Australia and The Hindu[7] from India. We compared the accuracy of the results obtained. These experiments and the results are presented in the following subsections.

4.1 Comparing Efficiencies of Four Types of NER Algorithms on a Regional Newspaper

Experimental set up - We collected 70 articles from Otago Daily Times that contained 'theft' as a search term using Mozenda. We tokenized these articles into sentences. We then investigated four different NER algorithms to find the one that yields the best results in identifying the locations correctly (step 3 of Figure 1).

The details of the four NER algorithms compared are given below.

1. *NLTK pre-trained named entity chunker -* The nltk named entity chunker (Bird et al. 2009) uses *ne_chunk* method from *nltk.chunk* module to identify the named entities such as person, organization, geo-political entites (e.g. city, region and country).
2. *Stanford NER -* Stanford Named Entity Recognizer is a Java-based tool (Finkel et al. 2005). A

[4] www.odt.co.nz
[5] www.nzherald.co.nz
[6] www.smh.com.au
[7] www.thehindu.com

No.	Features	Description	Sample keywords	Number of terms identified
1.	Crime Terms	Words or phrase that are related to theft crime	theft, burglary, stolen, failed to pay	336 (55)
2.	Police Locations	The word Police occurs after a specific location	Dunedin police, Timaru police	33
3.	Region Locations	The location identifies as a region	Otago, Canterbury	36
4.	City Locations	The location identifies as a city	Dunedin, Queenstown	122
5.	Suburb Locations	The location identifies as a suburb	Mary hill, Roslyn	88
6.	Street Locations	The location identifies as a street	Princes Street, Easther Cres	55
7.	Court Locations	The word Court occurs after a specific location	Dunedin District court	28
8.	Person Locations	Terms that describe a person occurs after a specific location	Dunedin man, Mosgiel woman, Wanaka pair, Auckland teenager	53 (19)

Table 2: Features defined and sample keywords

screenshot of the results obtained from this tool is given in Figure 4. It identifies four types of entities: location, organization, person and miscellaneous. We used the location information in this work. The algorithm uses the CRFclassifier to train the model for identifying named entities.

3. *NLTK chunker class using Gazetteer* - We used LocationChunker class from NLTK cookbook (Perkins 2010). It uses the gazetteers corpus to identify location words. Gazateer corpus is a large database containing the locations from all around the globe. However, the level of details available for each country varies. Users can upload their own (better) data to this corpus in order to obtain better results (i.e. reduce the number of false positives).

4. *LBJ Tagger* - LBJ NER Tagger (Ratinov & Roth 2009) is one of the models of the Named Entity Recognition system developed at the University of Illinois. This model is based on regularized average perceptron. It uses gazetteers extracted from Wikipedia, where the models for word class are derived from unlabelled text and expressive non-local features. We used the classic 4-label type model to identify the locations and organizations[8].

The results of these four algorithms to identify the locations correctly on the same set of 50 articles from Otago Daily Times is given in Table 4. Out of these four algorithms, LBJTagger has the highest accuracy (94%) followed by the NLTK Chunkparser (81%).

Algorithm	Precision	Recall	F-Score	Accuracy
NLTK pre-trained named entity chunker	0.93	0.78	0.85	0.74
StanfordNER	0.93	0.80	0.86	0.76
NLTK Chunkparser using Gazateer	0.88	0.91	0.90	0.81
LBJ Tagger	0.98	0.96	0.97	0.94

Table 4: Comparisons of four different NER algorithms based on location identification

[8]The data about organizations are used to in conjunction with the data about locations to in order to improve the accuracy of locations in our work.

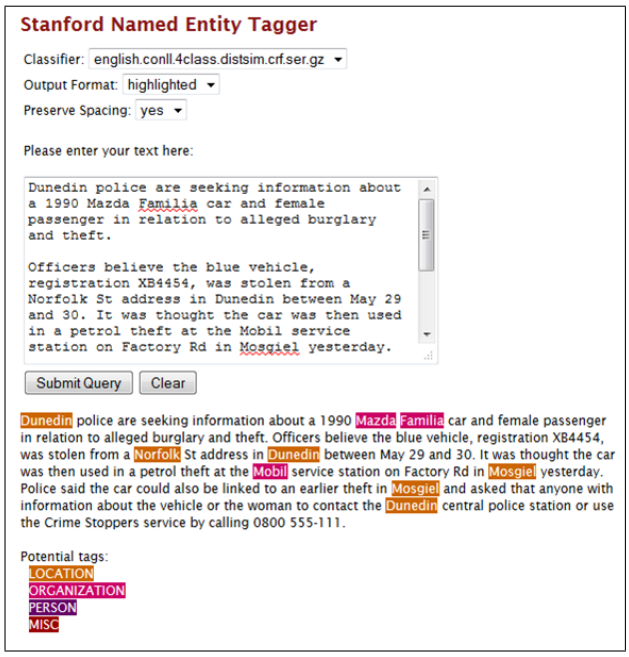

Figure 4: A sample snapshot of Stanford's online NER tool

4.2 Accuracy of Crime Sentence Labelling in a Regional Newspaper

We used the best NER algorithm, the LBJTagger to identify locations in the 70 Otago Daily Times (ODT) articles. Then, the features in all these articles were identified using a Python program which employed regular expressions (step 4 in Figure 1). The labels were assigned manually for the training set (step 5). We used Mallet, a Java-based tool (McCallum 2002), to train the CRF. Mallet uses SimpleTagger class for training and testing datasets.

We evaluated our results using 10-fold cross validation by splitting the dataset into 10 sets of training and test data sets. One set contained 63 training articles and 7 test articles (training to test ratio of 9:1 following the work of Angrosh et al. (2010)). Table 5 shows the precision, recall, f-score and accuracy for the test dataset. We achieved overall accuracy of (84% overall, with individual accuracies of CLS and

Table 5: Result of 10-fold cross validation (ODT articles using LBJ tagger)

Label	First and Zero order			
	Precision	Recall	F-score	Accuracy
CLS	0.87	0.88	0.87	0.77
NO-CLS	0.94	0.93	0.93	0.88

Table 6: Results of comparisons of three newspapers (First and Zero order)

Label	NZ Herald			
	Precision	Recall	F-score	Accuracy
CLS	0.90	0.94	0.92	0.85
NO-CLS	0.97	0.95	0.96	0.92

No-CLS being 77% and 88% respectively) using first and zero order CRF. The confusion matrix for the results obtained for the 70 articles (with a total of 523 sentences) is given in Figure 5. 155 sentences marked as CLS sentences by a human were also identified as CLS articles by the classifier model that has been developed in this work. 322 NO-CLS articles have also been identified correctly.

		Predicted class		
		CLS	NO-CLS	Total
Actual class	CLS	**155**	22	177
	NO-CLS	24	**322**	346
	Total	179	344	**523**

Figure 5: Confusion matrix for 70 articles from ODT for 10-fold cross validation

4.3 Comparison of Crime Location Extraction from Two Newspapers in New Zealand

Our hypothesis was that English used within New Zealand will be similar. Hence, the model that was created from training samples in Otago Daily Times (based in Dunedin) must be applicable for labelling articles of New Zealand Herald (based in Auckland). To test this hypothesis, we chose 50 theft-related articles from New Zealand Herald. The results obtained are presented in Table 6. It can be seen that we achieved a high accuracy (overall accuracy of 90% with individual accuracies for CLS and NO-CLS sentences being 85% and 92% respectively). So, the results obtained for New Zealand Herald is in support of our hypothesis that the model trained for ODT is applicable to New Zealand Herald. However, we need to conduct a large study involving more articles. Also, this might not be true to all countries. The use of English (e.g. written style) can vary from one part of the country to another.

4.4 Comparison of Crime Location Extraction Across Countries

We further hypothesized that the model developed for labelling newspaper articles in New Zealand might be adequate for classifying newspaper articles in Australia. In order to examine our hypothesis, we chose 50 theft-related articles from Sydney Morning Herald. The model reported in the previous step (i.e. model developed by training the 70 articles obtained from Otago Daily Times was our training data) was used to test the data from the 50 theft-related newspaper articles from Sydney Morning Herald. The results are presented in Table 7. It can be observed that the accuracy of the results was lower when compared to the artices from New Zealand newspapers (overall accuracy of 75% with individual accuracies for CLS and NO-CLS being 64% and 81% respectively). We also conducted a similar study on 50 articles from The Hindu. The accuracy was low (overall accuracy of 73% with individual accuracies for CLS and NO-CLS being 59% and 79% respectively).

Table 7: Results of comparisons of three newspapers (First and Zero order)

Label	Sydney Herald			
	Precision	Recall	F-score	Accuracy
CLS	0.69	0.90	0.78	0.64
NO-CLS	0.96	0.84	0.89	0.81

We investigated the reasons for difference in the accuracy. There were two main reasons. First, the efficiency of the LBJTagger on these articles was lower than the articles from New Zealand (i.e. locations were not identified correctly to start with which impacted the final results). Second, there were more instances of crime locations occurring in sentences that were apart (i.e. first sentence of the article talks about the crime and the fifth sentence specifies the crime location). Such instances were relatively rare in the articles obtained from New Zealand newspapers[9]. This primarily is an issue of writing style. Currently our work does not handle relationship between sentences. For example, it does not tie information from the first sentence (the crime information) and the fifth sentences (the location information) together. This is one of the areas for further work.

5 Discussion

We have have demonstrated the best results on sentence classification are obtained when LBJTagger is used. The accuracy of location identification is crucial for our approach, because the quality of this step (step 3) affects the subsequent steps. We have demonstrated that our approach works well for labelling sentences in Otago Daily Times articles into crime location sentences or not (accuracy of 84%). Also, we have demonstrated that re-usability of this model in the context of another newspaper (i.e. the accuracy for the articles in NZ Herald was 90%). However, the accuracies obtained by employing the same model for newspapers from other countries are slightly lower (75% and 73% for articles from Australia and India). We have discussed the underlying reasons and what needs to be done in the future. There are a couple of approaches in building models for crime sentence identification. The first one is to create individual models (one for each English speaking country). The second one is to create a global model which can be used for this purpose. The second one can be built from the datasets of the first one (i.e. global model can be built as an agglomeration of the local models).

We currently have extracted fine grained information from the theft related articles in Otago Daily Times and New Zealand Herald. These information include city location, suburb location. We used Google Fusion Tables (Halevy & Shapley 2009) to dis-

[9] This might not be the same for other crimes.

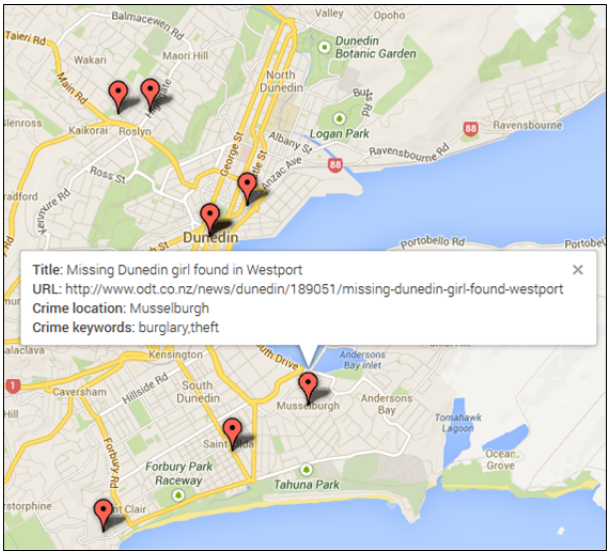

Figure 6: A snapshot of information displayed using the map view of Google's FusionTables

play this information. By clicking on a balloon on a map, the article related to that particular crime location can be viewed by the user (see snapshot shown in Figure 6). Currently, we provide details such as the title of the article, the URL, crime terms and the crime location. We are planning to modify this set up with a set of information that might be beneficial to a variety of stakeholders. For example, we plan to make the following pieces of information available to the user.

1. Offender's place of origin
2. Victim's place of origin
3. Police involvement details
4. Court involvement details
5. Involvement of organizations

There are a few limitations of the research work. First, the relationship between different sentences have not been explored. For example, the crime details may be in the first sentence and the location details can be in the fifth sentence of the same paragraph or even in the subsequent paragraph. This has not been modeled in this work. Using an appropriate relationship identification algorithm (e.g. Sutton & McCallum (2007)) for this purpose is the next step of this research. Second, we have not considered eliminating the duplicate articles reporting the same crime within a newspaper (e.g. elaborate news may follow brief news items) and across newspaper articles since our study was a feasibility study to demonstrate that our approach works. We plan to consider this in the future. Third, the approach uses a small sample size (70 articles in one newspaper) for the training data set. We believe, we will be able to improve the results by increasing the number of articles considered. Despite these limitations, we believe, the research reported in this work can be used to create a system which will be beneficial for visitors and immigrants to a city to make right decisions about where to stay/live and which areas to avoid. Also, the system will be useful for neighbourhood watch groups and city councils to monitor and prevent crimes.

A further extension of this study is to consider the full range of crimes as categorized in law (Australian Government 2013) and also extend this to other domains such as extracting historical record of cities on their cultural events, sports events, etc. Historical newspapers can be obtained from Archives New Zealand (New Zealand Government Archives 2013) that contain valuable historical events which can then be mined and visualized using a system like ours. For example, 19th century Dunedin can be visualized on top of the map based on the type of activities that were reported in newspaper articles between 1861 to 1900.

6 Conclusions

This paper presents a methodology for extracting crime location sentences (particularly 'theft' crime information) from online newspaper articles. It employs named entity recognition (NER) algorithms to identify locations in sentences and uses Conditional Random Field (CRF) to classify sentences into crime location sentences. The proposed system is evaluated on four newspaper articles from three different countries. It demonstrates that the accuracy of the results obtained for New Zealand articles varies from 84% to 90%. For articles from the two other countries (India and Australia) it varies from 73% to 75%. We have also discussed the limitations of our work and the areas for future improvements.

7 Acknowledgments

We would like to thank the Information Science department for the internal grant to support the first author to purse this research. Also, special thanks to Angrosh Mandya for his valuable help during the early stages of this work.

References

Angrosh, M., Cranefield, S. & Stanger, N. (2010), Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries, *in* 'Proceedings of the 10th annual joint conference on Digital libraries', ACM, pp. 293–302.

Australian Government (2013), '1234.0 - Australian and New Zealand Standard Offence Classification (ANZSOC), 2011', `http://www.abs.gov.au/ausstats/abs@.nsf/mf/1234.0/`. Accessed: 10-08-2013.

Bird, S., Klein, E. & Loper, E. (2009), *Natural language processing with Python*, O'reilly.

Chandra, B., Gupta, M. & Gupta, M. (2007), Adaptive query interface for mining crime data, *in* 'Databases in Networked Information Systems', Springer, pp. 285–296.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. & Chau, M. (2004), 'Crime data mining: a general framework and some examples', *Computer* **37**(4), 50–56.

Cowie, J. & Lehnert, W. (1996), 'Information extraction', *Communications of the ACM* **39**(1), 80–91.

De Bruin, J. S., Cocx, T. K., Kosters, W. A., Laros, J. F. & Kok, J. N. (2006), Data mining approaches to criminal career analysis, *in* 'Sixth International Conference on Data Mining 2006 (ICDM'06)', IEEE, pp. 171–177.

Finkel, J. R., Grenager, T. & Manning, C. (2005), Incorporating non-local information into information extraction systems by gibbs sampling, *in* 'Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics', Association for Computational Linguistics, pp. 363–370.

Furtado, V., Ayres, L., de Oliveira, M., Vasconcelos, E., Caminha, C., D'Orleans, J. & Belchior, M. (2010), 'Collective intelligence in law enforcement : The wikicrimes system', *Information Sciences* **180**(1), 4 – 17.

Halevy, A. & Shapley, R. (2009), 'Google fusion tables', `http://googleresearch.blogspot.co.nz/2009/06/google-fusion-tables.html`. Accessed: 25-08-2013.

Kiss, T. & Strunk, J. (2006), 'Unsupervised multilingual sentence boundary detection', *Computational Linguistics* **32**(4), 485–525.

Ku, C. H., Iriberri, A. & Leroy, G. (2008), Natural language processing and e-government: crime information extraction from heterogeneous data sources, *in* 'Proceedings of the 2008 international conference on digital government research', dg.o '08, Digital Government Society of North America, pp. 162–170.

Lafferty, J., McCallum, A. & Pereira, F. C. (2001), 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'.

McCallum, A. K. (2002), 'MALLET: A machine learning for language toolkit', `http://mallet.cs.umass.edu`. Accessed: 10-05-2012.

Mozenda (2013), 'Mozenda - a web data extractor tool', `http://www.mozenda.com`. Accessed: 10-05-2013.

Nadeau, D. & Sekine, S. (2007), 'A survey of named entity recognition and classification', *Lingvisticae Investigationes* **30**(1), 3–26.

New Zealand Government Archives (2013), 'Archives New Zealand', `http://archives.govt.nz`. Accessed: 10-08-2013.

NZ Government (2013), 'New Zealand Police Crime Statistics', `http://www.police.govt.nz/service/statistics/index.html`. Accessed: 15-08-2013.

Peng, F. & McCallum, A. (2006), 'Information extraction from research papers using conditional random fields', *Information Processing & Management* **42**(4), 963–979.

Perkins, J. (2010), *Python Text Processing with NLTK 2.0 Cookbook*, Packt Publishing Ltd.

Ratinov, L. & Roth, D. (2009), Design challenges and misconceptions in named entity recognition, *in* 'Proceedings of the Thirteenth Conference on Computational Natural Language Learning', Association for Computational Linguistics, pp. 147–155.

Sutton, C. & McCallum, A. (2007), 'An introduction to conditional random fields for relational learning', *Introduction to statistical relational learning* **93**, 142–146.

UK Police (2013), 'Local crime, policing and criminal justice website for england, wales and northern ireland', `http://www.police.uk`. Accessed: 15-08-2013.

Wiki Crimes (2013), 'Mapping crimes collaboratively', `http://www.wikicrimes.org`. Accessed: 15-08-2013.