Automatic Curation and Visualization of Crime Related Information from Incrementally Crawled Multi-source News Reports

Tirthankar Dasgupta, **Abir Naskar**, **Rupsa Saha** and **Lipika Dey**

TCS Innovation Lab, India
(dasgupta.tirthankar, abir.naskar, rupsa.s, lipika.dey)@tcs.com

Abstract

In this paper, we demonstrate a system for the automatic extraction and curation of crime-related information from multi-source digitally published News articles collected over a period of five years. We have leveraged the use of deep convolution recurrent neural network model to analyze crime articles to extract different crime related entities and events. The proposed methods are not restricted to detecting known crimes only but contribute actively towards maintaining an updated crime ontology. We have done experiments with a collection of 5000 crime-reporting News articles span over time, and multiple sources. The end-product of our experiments is a crime-register that contains details of crime committed across geographies and time. This register can be further utilized for analytical and reporting purposes.

1 Introduction

News articles from different sources regularly report crime incidents that contain details of crime, information about accused entities, details of the investigation process and finally details of judgment(Westphal, 2008). These details are not all published together, but pour in over time. A curated crime knowledge base with organized information about criminal activities and possible associated criminals is beneficial to a wide variety of end-users(Westphal, 2008; Furtado et al., 2010; Hassani et al., 2016; Arulanandam et al., 2014). While law-enforcers of a region have access to details of crime committed within their own jurisdiction, a shared knowledge-base containing information curated from open sources can be used by them to track criminal activities in other regions(Chau et al., 2002). These knowledge bases are also in demand by financial organizations who want to make use of these profiles to check on the credit-worthiness of a customer. Regulatory agencies also want to use these knowledge bases to verify legal compliance. Crime and corruption, common scourges of modern societies, top the list of problems cited by public entities in emerging and developing nations.

There are many challenges of such automatic information curation. Ensuring verifiability and reliability of information sources is a prime concern for curators. Handling factual variations or contradictions needs intelligent methods for disambiguation. Incremental compiling of facts from sources generated over a period of time also needs efficient entity resolution and linking mechanisms to ensure information continuity. While the present work addresses the later challenges, we do not delve deeper into the issue of information reliability but rather assume that News articles collected from reliable agencies supply authentic information.

The salient features of the demonstration are: a) We have leveraged deep convolution recurrent neural network to extract different crime related entities and events from News documents. This includes, *name of the accused, name of the victim, nature of crime, geographic location, date and time, law enforcement, charges bought, and action taken (if any) against the accused*. b) The model is trained over a dataset of 5000 documents that spans across time, different sources- containing a multiplicity of reporting on particular topics. c) We have done experiments with a collection of 1000 new News articles. The end-product of our experiments is a crime-register that contains details of crime committed across geographies and time. This register can be further utilized for analytical and reporting purposes.

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 103–107
Santa Fe, New Mexico, USA, August 20-26, 2018.

2 System Architecture

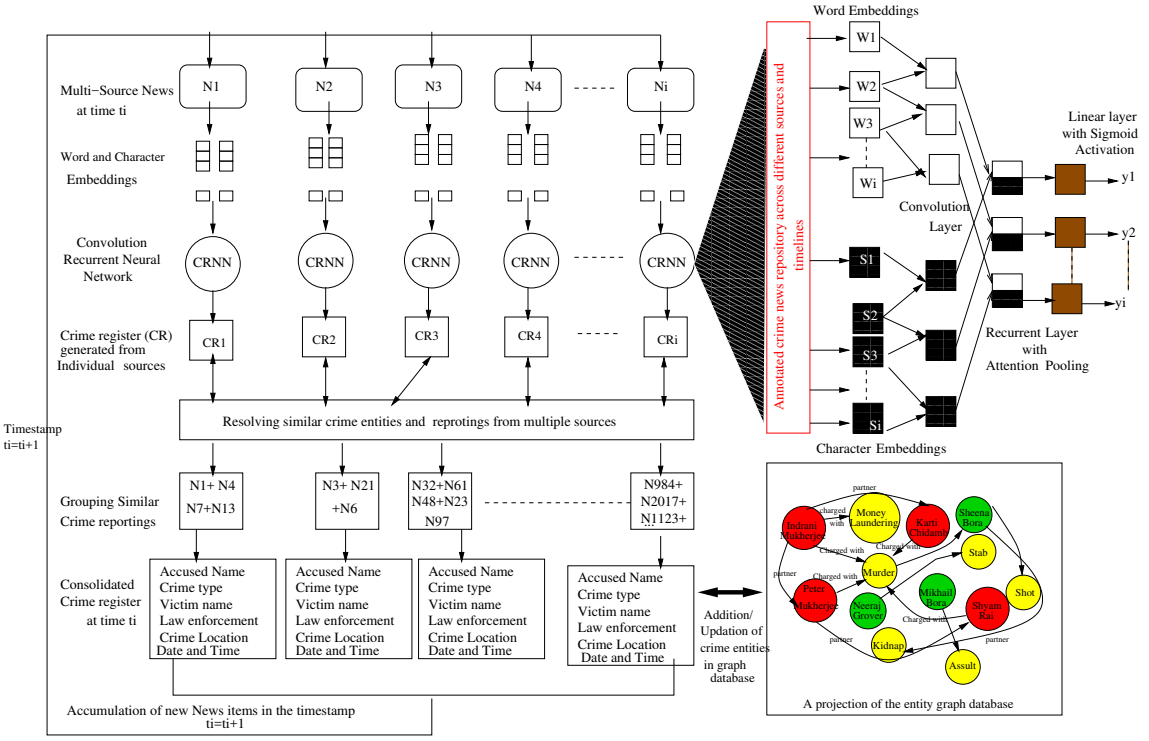

Figure 1: Overview of the crime event and entity extraction architecture.

The overall architecture depicted in Figure 1, consists of seven primary units: a) *Multi-source news accumulation*, b) Crime entity extraction using Convolution Recurrent Neural Network(CRNN) c) Creating *temporary crime registers*, d) *grouping similar crime reporting*, e) Updated *crime register*, and g) Reinforcement mechanism.

Multi-source news accumulation: The proposed system consists of multiple web crawlers that continuously crawls crime news events from 10 different sources in a near runtime manner.

Crime entity extraction using Convolution Recurrent Neural Network(CRNN): Output of the crawlers are passed to the crime entity extraction unit to extract the following crime components: *name of the accused, name of the victim, nature of crime, geographic location, date and time, law enforcement, charges bought, and action taken (if any) against the accused.* The input to the CRNN are a sequence of word and characters embeddings. We use the GloVe word vector representations of dimension 300 (Pennington et al., 2014). Both the word and character embeddings are passed to the convolution layer (CNN). CNNs works best in determining local features from texts. The CNN applies a linear transformation to all K windows in the given sequence of vectors. Given a word representations $X_1, X_2, ...X_l$, the convolution layer first concatenates these vectors to form a vector $\bar{x}$ of length $l.d_{LT}$ and then uses $Conv(\bar{x}) = W.\bar{(x)} + b$ to calculate the output vector of length d_c. Where, **W** and **b** are the weights that the network learns. The output of the CNN layer is passed to the RNN layer. We have used bidirectional LSTMs (Bi-LSTM) networks that are connected so that both future and past sequence context can be examined. After obtaining the intermediate layers from the Bi-LSTM, we use an attention pooling layer over the sentence representations. The attention pooling helps to acquire the weights of sentence contribution to final quality of the text. This is represented as: $a_i = \tanh(W_a.h_i + b_a)$, $\alpha_i = \frac{e^{w_\alpha.a_i}}{\sum e^{w_\alpha.a_i}}$, $O = \sum(\alpha_i.h_i)$. Where W_a, w_α are weight matrix and vector respectively, b_a is the bias vector, a_i is attention vector for i-th sentence, and α_i is the attention weight of i-th sentence. We apply a sigmoid function to limit the possible scores to the range $[0, 1]$. The mapping of the linear layer after applying the sigmoid activation function is given by $s(x) = sigmoid(w.x + b)$. Where, x is the input vector, w is the weight vector, and b is bias value.

Creating crime register for individual news sources: Once the target components are extracted, we create a temporary crime register of crime entities extracted from each of the individual source news items. Figure 2 and Figure 3 illustrates the basic working of the crime extraction tool. In the left panel of Figure 2, there are series of news headline corresponding to a particular date. Once a headline is selected, the corresponding news details are displayed in the middle panel. Corresponding to the detailed news all the respective crime entities and events were extracted and labeled in different colors. There are number of instances where the same crime event is reported in multiple news sources in different ways. There are also issues related to information richness of one source as compared to other. In particular we have observed that regional news sources cover deeper information rather than nationalized news sources. In addition to this there are severe issues related to entity resolution both within the document as well as outside the document. For example, "*Sailesh Patel*" may be referred as "*Sailesh*" or "*Patel*" in other articles. In order to resolve such entities, we have used the BMI measure as discussed in (Lau et al., 2008; Lau et al., 2009). Also, resolution of such named entities help grouping similar crime reporting together. Thus, it is also important to ensemble information across different news sources together and construct a unified knowledge representation.

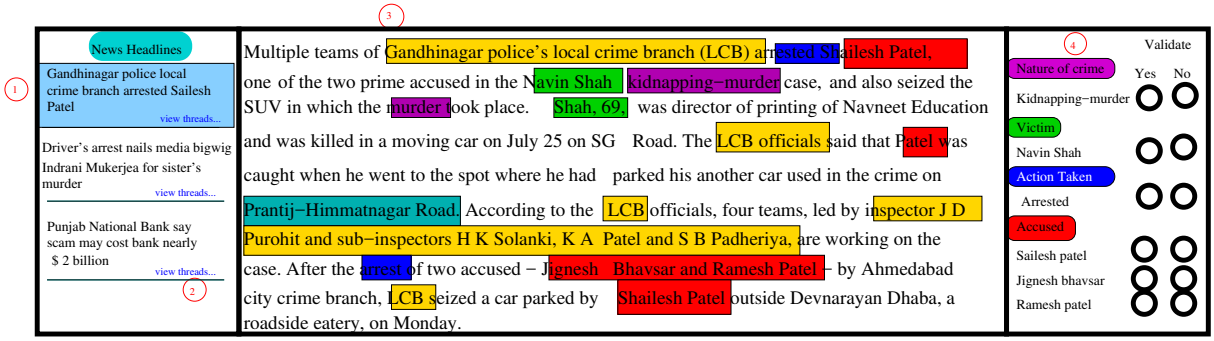

Figure 2: Working of the crime extraction module. The extracted labels are marked in color.

Tracking new progress and outcomes: As discussed earlier, all the details of a crime incident are not published together, but gradually reported over time. This may span between days to months to year. Over the time information regarding the crime incident changes and new outcomes emerges. For example, in Figure 2 our system demonstrates how a particular crime incident about "Sheena Bora murder case" changes over time with new accused names, crime type and victims. Therefore, it becomes extremely difficult for a curator to manually keep track of all the records form the past repositories. With respect to this, the present system plays an important role in automatically identifying, tracking, and maintaining crime incidents that last over years. Once the crime entities of similar crime reports are grouped together, the entities along with their relations are then stored in an crime ontological structure. If a new entity arrives, then the corresponding ontology along with its relations will be updated in the repository.

As discussed earlier, the proposed system in mainly developed to assist analysts and knowledge workers for exploring, reviewing and visualize textual data. With respect to this an important feature of the proposed system is its ability to adapt based on the users feedback. For example, in Figure 2 the system provides option to the user to change the system predicted outputs. Based on the users output, the system has the capability to retrain its classification and extraction model. This in turn help create new and enriched models.

2.1 Experimentation and Evaluation

In order to evaluate, we took another set of 1000 news articles from 3 different sources over the period of five years. Each of the articles were manually annotated by a group of experts. The annotation process involves identifying the major crime components as discussed earlier. The system is evaluated by comparing its output with that of the expert annotations. We quantify the performance score in terms of the precision(P), recall(Re), and F-measure(F). Table 1 reports the comparison of performance of

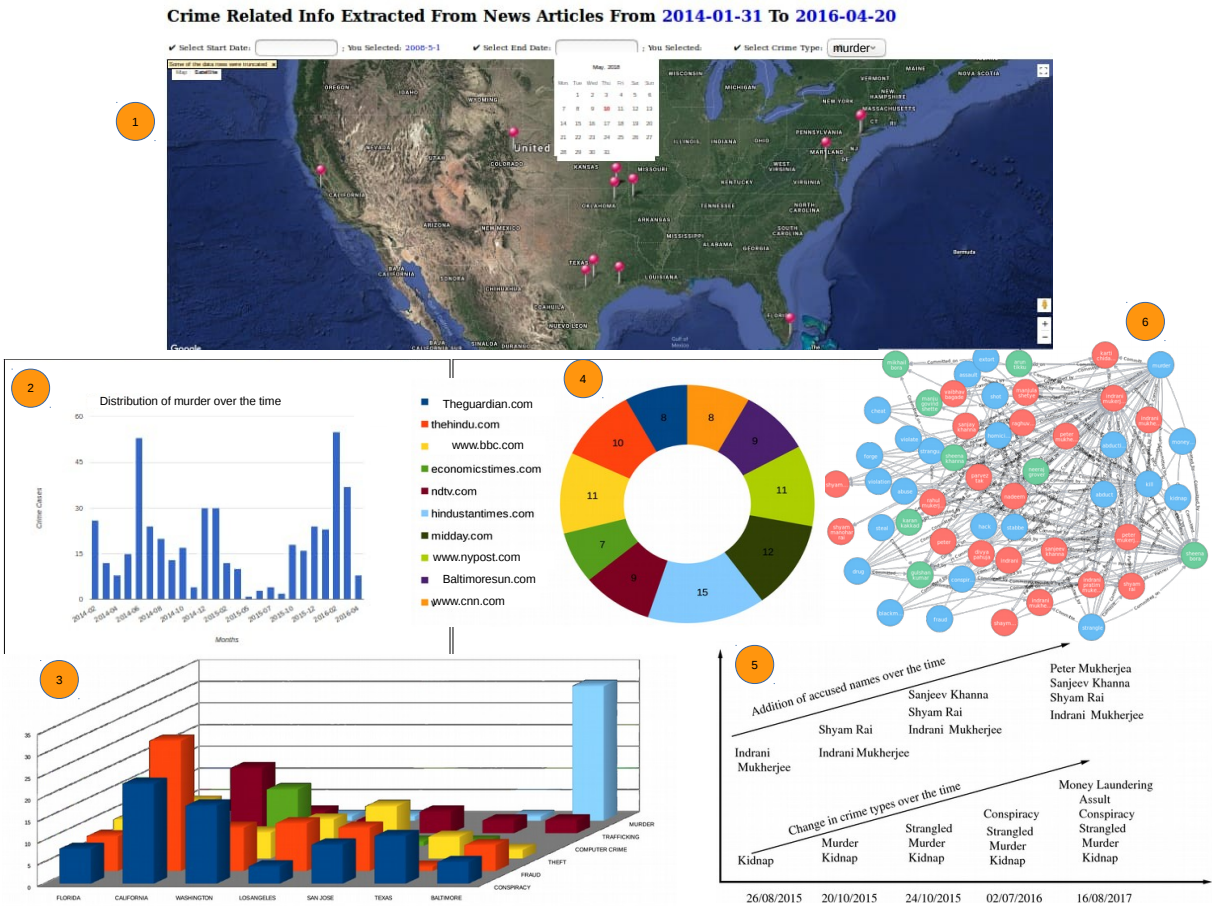

Figure 3: Working of the visualization tool. (1) display a map view that projects distribution of crime across geographical regions. (2) displays distribution of a particular crime over the years (3) shows distribution of crime over geographic region. (5) shows an illustration of a particular news where how crime information changes over time. (6) displays a projection of crime entity knowledge graph constructed from the raw news documents over the past five years.

Table 1: Comparing the F-measure score of different crime entity extraction models.

	Accused	Crime type	Location	Victim	law enforcement	Action taken	Charges
CNN	69	74	68	58	69	59	73
RNN	73	72	63	67	55	62	76
CRNN	77	71	73	67	72	68	81

the CRNN model with respect to other baseline models like, CNN based model, RNN model and our proposed CRNN model. For the sake of space we have reported only the F-measure scores.

References

Rexy Arulanandam, Bastin Tony Roy Savarimuthu, and Maryam A Purvis. 2014. Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference-Volume 155*, pages 31–38. Australian Computer Society, Inc.

Michael Chau, Jennifer J Xu, and Hsinchun Chen. 2002. Extracting meaningful entities from police narrative reports. In *Proceedings of the 2002 annual national conference on Digital government research*, pages 1–5. Digital Government Society of North America.

Vasco Furtado, Leonardo Ayres, Marcos De Oliveira, Eurico Vasconcelos, Carlos Caminha, Johnatas DOrleans, and Mairon Belchior. 2010. Collective intelligence in law enforcement–the wikicrimes system. *Information Sciences*, 180(1):4–17.

Hossein Hassani, Xu Huang, Emmanuel S Silva, and Mansi Ghodsi. 2016. A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(3):139–154.

Raymond YK Lau, Peter D Bruza, and Dawei Song. 2008. Towards a belief-revision-based adaptive and context-sensitive information retrieval system. *ACM Transactions on Information Systems (TOIS)*, 26(2):8.

Raymond YK Lau, Dawei Song, Yuefeng Li, Terence CH Cheung, and Jin-Xing Hao. 2009. Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE transactions on knowledge and data engineering*, 21(6):800–813.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Christopher Westphal. 2008. *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press.