Named Entity Recognition in Crime Using Machine Learning Approach

Hafedh Shabat, Nazlia Omar, and Khmael Rahem

Knowledge Technology Group, Center for AI Technology, Faculty of Information Science and Technology, Universiti Kebangasaan Malaysia, 43600 Bangi, Selangor, Malaysia
h2005_ali@yahoo.com, nazlia@ukm.edu.my, khm2006_rakem@yahoo.com

Abstract. Most of the crimes committed today are reported on the Internet by news articles, blogs and social networking sites. With the increasing volume of crime information available on the web, a means to retrieve and exploit them and provide insight into the criminal behavior and networks must be determined to fight crime more efficiently and effectively. We believe that an electronic system must be designed for crime named entity recognition from the newspaper articles. Thus, this study designs and develops a crime named entity recognition based on machine learning approaches that extract nationalities, weapons, and crime locations in online crime documents. This study also collected a new corpus of crime and manually labeled them. A machine learning classification framework is proposed based on Naïve Bayes and SVM model in extracting nationalities, weapons, and crime location from online crime documents. To evaluate our model, a manually annotated data set was used, which was then validated by experiments. The results of the experiments showed that the developed techniques are promising.

Keywords: Crime, Machine Learning, Named Entity Recognition, Support Vector Machine, Naïve Bayes.

1 Introduction

With the rapid growth of the World Wide Web, the volume of crime information available is growing exponentially. Media channels, especially print and broadcast media, have been replete with reports of crime and violence. As a result, the process of manually analyzing and processing the extensive data has become a difficult task. The public, journalists, crime analysts, and policemen depend on resources reported by media in investigating or monitoring crime. Finding relevant and timely information from crime documents are crucial for many applications and can play a central role in improving crime-fighting capabilities, enhancing public safety, and in reducing future crimes. Useful and highly important information and entities, such as the victims and their names, names of organizations, crime locations, used weapons, and crime types, should be extracted. The crime domain has been chosen as an area in this study because of its social importance. Named entity recognition system must be

A. Jaafar et al. (Eds.): AIRS 2014, LNCS 8870, pp. 280–288, 2014.

developed to process crime news and topics and detect new reports on a specific crime type. This system is beneficial to users, such as the public, in obtaining awareness on crimes committed in the past and at present.

This study aims to propose NER model that utilizes two techniques (Support Vector Machine (SVM) and Naïve Bayes(NB)) to extract data including weapons, crime locations, and nationalities of both the victim and the suspect. This model is targeted to be more effective than previously developed models.

2 Related Work

In the recent decade, several studies have been performed on crime data mining. The results are often used in developing new software applications for detecting and analyzing crime data. [1] designed neural network-based entity extraction techniques to recognize and classify useful entities from police narrative reports. The designed named entity recognition (NER) has five entity identifier types that crime investigators believe would be useful in their crime investigations. The five types are "person", "address", "vehicle", "narcotic drug", and "personal property". Result showed that the system achieved encouraging precision recall rates for person names and narcotic drugs, but did not perform well for addresses and personal properties. The limitation of this study is that only two named entities, namely, person name and drugs, were extracted, unlike the present study that will extract weapons, type of crime, location of crime, nationality of the victim, nationality of the suspect, and the association between the weapons and the type of crime.

[2] described a rule-based prototype for identifying types of crime in a text in the crime domain, which utilizes a list of keywords and a matching algorithm for classification. [3] had worked on addressing the crime pattern detection problem by using a clustering algorithm to help detect crime patterns and to hasten the process of solving crimes. The k-means clustering technique was employed with some improvements to support the process of identification of crime patterns in actual crime data acquired from the office of a sheriff. Result showed that the most important attribute in crime patterns are "race", "age", "gender", and "weapon". [4] combined IE and principles of the cognitive interview. The IE system combined a large crime-specific lexicon, several General Architecture for Text Engineering (GATE) modules, and an algorithm to recognize relevant entities. The cognitive interview is a psychological technique in which people recall information about an incident. Commonly extracted crime-related entities are race, gender, age, weapons, addresses, narcotic drugs, vehicles, and personal properties. In this paper, crime entities are combined to 15 categories, namely, "act/event", "scene", "people", "personal property", "vehicle", "weapon", "body part", "time", "drug", "shoes", "electronic", "physical feature", "physical condition", "hair", and "clothing". Result showed high precision recall for narrative types. [5] also created another similar system that recognized crime-related information besides the type of crime. The newer system has the capability to extract the nationalities of the victim and the

suspect as well as the location of the crime, aside from the crime type that can be extracted by the previous model. The system utilized an indicator in the Arabic language, which did not work correctly for English because the English language has its own multiple indicator terms. [6] suggested a model that relies on natural language processing methods and adopted the Semantic Inferential Model. The model utilized collaborative environments on the Internet. The system is called Wiki Crimes and can extract two main crime entities, namely, crime scene and crime type from online web sites. [7] developed an IE model that focused on the extraction of information specific to theft crimes, specifically the crime location (address), from newspaper articles. Theft information was extracted from newspaper articles in three different countries: New Zealand, Australia, and India. The model implemented entity recognition to reveal if the sentence indicated a crime location or not and the conditional random field approach, which is a machine learning method, to check whether or not a sentence shows the crime location.

3 Methodology

In Constructing a system for NER using a machine learning approach requires many computational steps, including data planning, pre-processing, feature extraction, classification, and evaluation. Fig. 1 shows the overall architecture of the method, which involves the following phases:

1. Language resource construction phase
2. Pre-processing phase
3. Feature extraction phase
4. Crime entity extraction and classification phase
5. Evaluation phase

The following subsection will describe each of the phases.

3.1 Language Resource Construction

The use of a supervised machine learning technique relies on the existence of annotated training data. Such data are usually created manually by humans or experts in a relevant field. The data used in this research were collected from the Malaysian National News Agency (BERNAMA). The weapons, locations, and nationalities mentioned in the documents were annotated manually. Each file consisted of an article by a journalist reporting on one or more crimes.

3.2 Pre-processing Phase

The pre-processing phase is a highly important phase in any system that uses a machine learning framework. Before extraction and classification, each document undergoes the pre-processing phase that has the following steps:

1. Tokenizing of words depending on the white space and punctuation marks
2. Stop word removal: Removing punctuation marks, diacritics, non-letters, and stop words.
3. Part of Speech (POS) disambiguation is the ability to computationally determine the POS of a word activated by its use in a particular context. In this step, each word is tagged using its unique POS tag. Fig. 2 shows a sample of a crime text annotated with POS tags.

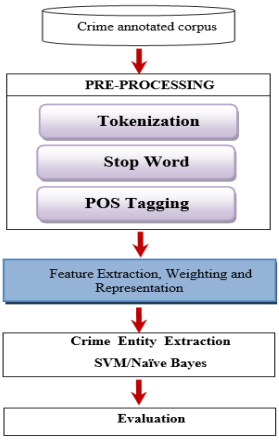

Fig. 1. The proposed crime entity extraction system architecture

On/IN the/DT first/JJ count/NN Nantha/NNP Kumar/NNP was/VBD charged/VBN with/IN a/DT few/JJ others/NNS who/WP are/VBP still/RB at/IN large/JJ when/WRB armed/VBN with/IN a/DT knife/NN to/TO have/VB robbed/VBN R/NNP Victor/NNP Ratnum/NNP 19/CD of/IN RM350/CD a/DT handphone/NN two/CD gold/NN rings/NNS a/DT silver/NN

Fig. 2. Sample of a crime text annotated with POS tags

3.3 Feature Extraction

Feature extraction is an important step in any classification system because it improves the efficiency of the classification tasks in terms of speed and effectiveness of learning. This phase aims to convert each word to a vector of feature values. This study defines a set of features for extracting information on nationalities, weapons, and crime locations from online crime documents. These features are grouped into three main feature sets: features based on POS tagging, features based on word affixes, and features based on the context. Table 1 details these feature sets. The word in the corpus is represented using the following feature vector.

Table1. Summary of the feature sets

Feature Set Name	Feature	Feature Name
Word affixes	F1	Prefix1
	F2	Prefix2
	F3	Prefix3
	F4	Suffix1
	F5	Suffix2
	F6	Suffix3
Context-based features	F7	Is the first letter of the word capitalized?
	F8	Previous word (window size 2)
	F9	Next words (window size 2)
	F10	Number of indicator words before (size of window 7)
	F11	Number of indicator words after (size of window 7)
	F12	Distance between previous indicator words and the current words
	F13	Distance between the resulting indicator words and the current words.
POS-based	F14	Part of speech of the previous three words
	F15	Part of speech of the next three words
	F16	Is the word part of the previous phrase?

3.4 Machine Learning and Classification

Most of the machine learning approaches has two phases in which training is first performed to generate a trained machine and then followed by a classification step. In this study, we will evaluate specific machine learning approaches. However, to extract the crime entities, namely, nationalities, weapons, and crime locations from online crime documents, the following machine learning classifiers are used:

SVM Classifier. Support vector machine (SVM) is an effective machine learning technique first introduced by [8]. SVMs are popular in the machine learning community because of their use of text categorization. The SVM classification method has been proven by many researchers as one of the most effective classification methods based on its performance on text classification and entities recognition [9-13]. Adopting the structural risk minimization principle from the computational learning theory, SVMs seek a decision surface to separate the training data points into two classes and to make decisions based on the support vectors selected as the only effective elements in the training set. Multiple variants of SVMs have been developed [10]. In this paper, our discussion is limited to linear SVMs because of their popularity and high performance. The optimization procedure of SVMs (dual form) is aimed to minimize the following:

$$\vec{\alpha} = argmin\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \overrightarrow{x_i}, \overrightarrow{x_j}\rangle\} \quad (1)$$

$$Subjectto: \sum_{i=1}^{n}\alpha_i y_i = 0; \quad 0 \le \alpha_i \le C$$

Naive Bayes. The Naïve Bayes (NB) algorithm is a widely used algorithm for review classification. The main advantages of NB are that they are simple, easy to implement, and comprise better-performing algorithms. Given a feature vector table, the algorithm computes the posterior probability that the review belongs to different classes and assigns the review to the class that has the highest posterior probability. Two models (i.e., the multinomial model and the multivariate Bernoulli model) are commonly used in the application of the NB approach to text categorization. NB assumes a stochastic model of document generation and implements the Bayes' rule. To classify the most probable class c* for a new document d, NB computes:

$$c^* = \arg\max_c \; P(c/d) \quad (2)$$

The NB classifier calculates the posterior probability as follows:

$$p(c_j|d_i) = \frac{p(c_j)p(c_j|d_i)}{p(d_i)} \quad (3)$$

4 Performance Measures

The performance of the machine learning algorithms is measured on manually labeled crime corpus. The corpus contains 500 documents collected from the Malaysian National News Agency (BERNAMA). Each file consisted of an article by a journalist reporting on one or more crimes. The weapons, locations, and nationalities mentioned in the documents were annotated manually. Fig. 3 show a sample of the annotated of the used data set.

…….. with a hold-up at a convenience store in Jalan\LOC-I Kuantan-Kemaman\ LOC–O yesterday ……The police recovered loot comprising RM2 500 homemade\WT-I pistol\WT-O two shotgun\WT-O pellets several knives\WT-O and a Perodua Myvi from the trio The suspects had taken away the car after robbing the store which was manned by two workers One of the suspects is believed to have supplied the gun\WT-O

Fig. 3. A Sample of the used dataset

All algorithms are evaluated using 10-fold cross-validation. The objective of this step-up is to filter the parameters and select the best methods for crime entity extraction. The performance measures used to evaluate the named entity recognition systems participating in the CoNLL-02, CoNLL-03, and JNLPBA-04 challenge tasks are precision, recall, and the weighted mean Fβ=1-score. Precision is the percentage

of named entities found by the learning system as correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if an exact match of a corresponding entity in the data file, that is, the complete name of the entity, is identified. Definitions of the performance measures used are summarized below. The same performance measures are employed to evaluate the results of the baseline experiments.

$$\text{Recall} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities in the corpus}} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{4}$$

$$\text{Precision} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities found by algorithm}} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{5}$$

$$F_{\beta} = \frac{(1+\beta^2) * (\text{precision} * \text{recall})}{(\beta^2 * \text{precision} + \text{recall})} \tag{6}$$

5 Results

In this experiment, the overall performance of each of two classifiers on crime entity extraction was examined. The two classifiers, NB and SVM, are applied on the entire feature space. Table 2 shows a summary of the experimental results using the NB and SVM classifiers in extracting the nationalities of the victim and the suspect, weapons, and crime locations.

Table. 2. Performance (the average F-measure for each class) of the NB and SVM classifiers on each crime entity

	SVM	NB
Weapons	91.08	86.73
Nationality	96. 25	94. 02
Location	89.28	87.66

As shown in Table 2, the highest performance on extracting weapons, nationalities, and crime locations from crime documents is exhibited by the SVM classifier. SVM uses a refined structure that acknowledges the relevance of most features. SVM also has the ability to handle large feature spaces. Therefore, probability is a good choice in the crime data set as the feature set cases of each example frequently occur.

The experiments in this study have generally shown highly promising results that clearly demonstrate the appropriateness of the application of machine learning algorithms for crime entity extraction. The promising result encourages the more

comprehensive and comparative study of machine learning for all crime entity extraction techniques.

6 Conclusion

This paper presented a crime NER based on machine learning classification approaches, namely, NB and SVM, in extracting nationalities, weapons, and crime locations from online crime documents. This study collected a new corpus of crime and manually labeled them. A manually annotated crime dataset was used for the evaluation, which was then validated by experiments. Finally, the results demonstrated that the SVM classifier achieves the best performance and outperforms NB in extracting nationalities, weapons, and crime locations from online crime documents.

Future research may be targeted at developing a large crime corpus and designing a general framework for crime IE and analysis, which includes automatic classification of crime type, identification of weapons used in each crime, identification of nationalities of both victim and suspect, identification of crime location, and analysis on the association between the used weapons and the type of crime.

References

1. Chau, M., Xu, J.J., Chen, H.: Extracting Meaningful Entities from Police Narrative Reports. In: 2002 Proceedings of the 2002 Annual National Conference on Digital Government Research, pp. 1–5 (2002)
2. Alruily, M., Ayesh, A., Al-Marghilani, A.: Using Self Organizing Map to Cluster Arabic crime documents. In: Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT, pp. 357–363 (2010)
3. Nath, S.V.: Crime Pattern Detection using Data Mining. In: 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2006 Workshops, pp. 41–44. IEEE (2006)
4. Chih Hao, K., Iriberri, A., Leroy, G.: Crime Information Extraction from Police and Witness Narrative Reports. In: Conference on Technologies for Homeland Security, pp. 193–198. IEEE (2008)
5. Alruily, M., Ayesh&, A., Zedan, H.: Automated Dictionary Construction from Arabic Corpus for Meaningful Crime Information Extraction and Document Classification, 137–142 (2010)
6. Pinheiro, V., Furtado, V., Pequeno, T., Nogueira, D.: Natural Language Processing based on Semantic Inferentialism for Extracting Crime Information from Text. In: IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 19–24 (2010)
7. Arulanandam, R., Savarimuthu, B.T.R., Purvis, M.A.: Extracting Crime Information from Online Newspaper Articles. In: Proceedings of the Second Australasian Web Conference (2014)
8. Cortes, C., Vapnik, V.: Support-vector Networks. Machine Learning 20, 273–297 (1995)
9. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization (1997)
10. Joachims, T.: The Maximum Margin Approach to Learning Text Classifiers: Methods,Theory, and Algorithms. PhD thesis, university Dortmund (2001)

11. Joachims, T.: Text Categorization With Support Vector Machines: Learning with Many Relevant Features. In: European Conference on Machine Learning, Chemnitz, Germany, pp. 137–142 (1998)
12. Isa, D., Lee, L.H., Kallimani, V.P., RajKumar, R.: Text Documents Preprocessing with the Bahes Formula for Classification using the Support Vector Machine. IEEE, TKDE 20(9), 1264–1272 (2008)
13. Saha, S., Ekbal, A.: Combining Multiple Classifiers using Vote based Classifier Ensemble Technique for Named Entity Recognition. Data& Knowledge Engineering, 85 (2013)