Review

JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

Vol.32; No.1.| 2024

Information Extraction Models for Crime Domain: A Review

Muayad N. Abdullah[1*]
1College of IT, University of Babylon, muayad.masudi@uobabylon.edu.iq
***Corresponding author email:** muayad.masudi@uobabylon.edu.iq; mobile: 07811583188

نماذج استخلاص المعلومات في مجال الجريمة: مراجعة

مؤيد نجم عبدالله[1*]
1 كلية تكنولوجيا المعلومات، جامعة بابل، muayad.masudi@uobabylon.edu.iq

Accepted: **31/12/2023** **Published:** **31/3/2024**

ABSTRACT

Information Extraction refers to the systematic technique of obtaining valuable information from unstructured texts. Named Entity Recognition is a part of information extraction, and is the process of extracting named entities of interest from text, such as person names, places, events, and entities, among others. Named Entity models are useful in critical situations where time and accuracy are important, such as when crime analysts need particular information about a crime case to help them solve it. In the crime domain, police and crime analysts need immediate information on certain crime cases to solve the crime or prevent it from happening again. Crime news documents consist of details on crimes and these details make these documents beneficial. crime analysts are able to quickly and accurately extract beneficial information from unstructured text. With the increase in crime rates all over the world because of the increase in the population, many Named Entity models have been proposed to help crime analysts solve crime cases. Many studies have been conducted on the performance of these methods based on general News wire articles. This work reviews several Named Entity models used for the crime domain.

Key words: information extraction, named entity, machine learning, rule-based, crime domain.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

1. INTRODUCTION

The widespread availability of the internet has facilitated real-time access to global news for individuals worldwide. Social media has enabled the rapid circulation of news, allowing content to be easily shared with friends or followers. As a result, information spreads quickly across social networks. [1].

Information Extraction (IE) deals with the automatic extraction of useful information from unstructured text documents [2,3]. Named Entity Recognition (NER) is used to extract useful entities from text. These NE models are used in order to make it easy for people to quickly and accurately extract needed information in order to satisfy an information need. This is especially useful in critical situations where time and accuracy is important, such as when crime analysts need particular information about a crime case to help them solve it. With the increase in crime rates all over the world because of the increase in the population [4,5], more research has been seen to create reliable NE models for the crime domain to help analysts with crime cases. NER identifies, categorizes and extracts the most important pieces of information from unstructured text without requiring time-consuming human analysis. It's particularly useful for quickly extracting key information from large amounts of data because it automates the extraction process. This review focuses on NER models for the crime domain. The different techniques used by each model are considered (supervised machine learning, rule-based, etc.), and also the information that each model extracts (suspect name, crime location, weapons, drugs, type of crime, etc.) [6].

2. INFORMATION EXTRACTION

In general, NE models heavily rely on techniques from NLP and IR. This allows them to retrieve specific and important information that is inside text of unstructured nature, such as text documents, news articles, among others. NE does not require to have a great comprehension of the actual meanings of the text. The information that is extracted and retrieved may be shown to the user, or also used by analysts for analyzing certain information [7].

The main difference between IR and IE is that IE only retrieves that particular information needed, or features, rather than the entire text document. IR retrieves entire text documents. For example, a user doing a Web search on Google and gets back several relevant document is an example of IR. With IE (particularly NE), on the other hand, is shown in the example of a user that aims to extract all of the drug names from a crime news article.

Information Extraction (IE) is consists of generally five main subsections. Every section deals with a particular type of extraction method [7]. All of the methods are named entity extraction (NE), template element (TE), co-reference resolution (CO), scenario template production (ST), and template relation (TR).

The first technique, which is NE, is the easiest to use, and has high accuracy. It is easiest to use since in this technique, it attempts to recognize and extraction named entities primarily, for example, people, places, organizations, objects, among others. NE models are capable of extracting information with an overall accuracy of 95%. This technique is almost near the capability of humans. Another advantage of this technique is that it is not dependent on any particular domains. Instead, it may work well on any domain [8].

CO tries to recognize if there are any certain relationships between two different words or terms in a text. There are two main issues in this technique, which are anaphoric resolution and proper noun resolution. Also, this technique, unlike NE, is depends on the domain that is being used, and whatever the domain used, a knowledge base that is related to the domain must be incorporated. CO is about 60% effective [9].

TE is relies on the overall outcomes of both NE and CO, and links text information with any extracted entities. TR needs the recognition of a limited amount of potential links among all template components that have been recognized in the TE. The ST models are the remaining section of IE models, and they bring jointly all information on TE entities and TR relations to generate event characterizations. This study will take into consideration NE and review the latest NE models for the crime domain.

3. INFORMATION EXTRACTION ARCHITECTURE

In general, an information extraction system receives an input, which is usually a text document, article, or web page. After that, the input is taken to the preprocessor for removal of the stop words, punctuation, or html/xml tags, and the tokenization occurs. Next, the extraction process is conducted, and relies on the extraction methods that are used by the model, such as rule based, lexical lookup, or statistical, among other approaches. The extracted information is sometimes sent to a template or database, depending on the steps that must be taken in the model. Figure 1 presents a very high level of a typical architecture of an information extraction system [10].

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

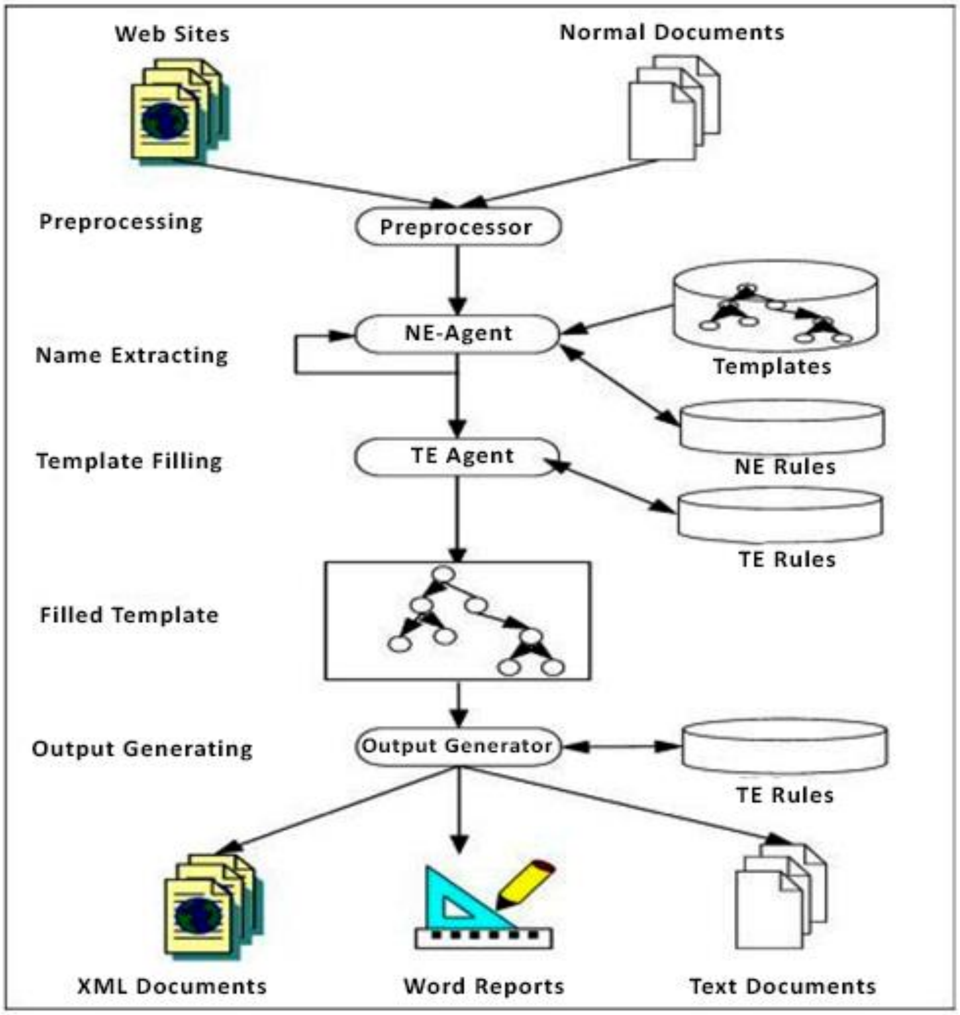


Figure 1. General information extraction system architecture [10]

4. TECHNIQUES USED IN NAMED ENTITY RECOGNITION

The techniques for Named Entity Recognition (NER) can be broadly categorized into three groups: rule-based approaches, learning-based techniques, and hybrid methods [11].

4.1 Rule-based Methods

Prior NER systems predominantly relied on manually designed rules. These systems employed information lists, such as gazetteers (books containing detailed lists of places), and rule-based syntactic-lexical patterns to identify and classify items [12]. By employing specific linguistic domain features, they get significant accuracy. The rule-based approach has a number of drawbacks, such as domain-specificity, expensive creation and maintenance costs, and limited mobility. Moreover, they require human competence in language comprehension and

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

programming skills. Because of this, rule-based techniques are inapplicable to many language domains, which has academics concentrating on machine learning techniques. Articles like the ones listed use rule-based approaches. It is empty in the user's text [13]. The Bengali language utilizes a rule-based approach, which poses challenges in sustaining Named Entity Recognition (NER) due to its language-specific nature. Rule-based learning is characterized by a limited number of elements that adhere to distinct patterns, such as dates, emails, and time. The rule-based technique has superior accuracy in identifying items with discernible patterns compared to the machine-learning method. The rule-based approach faces difficulties due to its language dependence.

4.2 Machine Learning Methods

The objective of a named entity recognition framework based on machine learning is to convert identification difficulties into classification problems. The problem is then solved using a statistical model [14]. Named Entity Recognition (NER) employs machine learning techniques to analyze text and identify connections and patterns. Machine learning algorithms offer several advantages compared to rule-based systems. They possess the ability to be trained, can easily adapt to different domains, and require lower maintenance costs for trained data. Cutting-edge NER often employs machine learning methods that rely on statistical machine learning. Machine learning (ML) techniques can be categorized into two main types: supervised learning, and unsupervised learning [15].

4.2.1 Supervised Machine Learning

Supervised learning, often known as classification, is a machine learning method that involves utilizing a function to accurately classify or categorize data. The training data comprises a collection of training data. In supervised learning, each data sample comprises an input entity, often represented as a vector, and a corresponding output estimate, which serves as the supervisory signal. A supervised learning operation analyzes the training data and generates a model that can be applied to future data samples.

An effective case enables the operation to accurately classify new inputs. This necessitates the learning process to effectively handle unfamiliar input from the training data. In order to address this difficulty in supervised learning, the following stages must be carried out:

1. Recognize the type of training data samples. First, the user has to select the kind of data that is to be used as the training data set.
2. Gather an entire training set. The training set needs to be similar to the actual implementation of the existing function. Consequently, a set of input entities is gathered and related outputs are gathered also, from sources of measurements.
3. Recognize the input feature symbol of the existing function that is trained. The accuracy of the learned function depends greatly on the way the input entity appears. Usually, the

input entity is transformed into a certain feature vector that consists of multiple features that supply information on the entity. The number of features has to be limited, because of the dimensionality issue, and also include enough information to accurately predict the output at the same time.

4. Recognize the general layout of the learned function and related learning algorithm. For example, the researcher may choose to use support vector machines (SVM).
5. Conduct the implementation. Run the learning algorithm on the gathered training set. Multiple supervised learning algorithms require the application user to recognize certain control configurations. These configurations are adjusted by increasing the efficiency on a set entitled the validation set of the training set, or via a cross-validation technique.
6. Evaluate the accuracy of the learned function. As soon as the configurations are tuned and learning is conducted, the performance of the derived function has to be estimated on a test set that varies from the training set that was implemented.

This review demonstrates three primary methodologies in supervised learning for extracting valuable crime-related information from text. The techniques are as follows:

A- Support Vector Machines (SVM)

SVM is a sort of machine learning technique that requires labeled data for training. Unlike neural networks, which only seek to identify a dividing hyperplane within a given instance, Support Vector Machines (SVM) aim to locate the optimal hyperplane within a fed space. The SVM technique in NER identifies four types of named entities: person, location, organization, and miscellaneous, utilizing Hindi and Bengali languages [16]. The findings of the Hindi experiment yielded a precision rate of 90.22%, a recall rate of 89.41%, and an F-score of 89.81%. The precision of Bengali was 91.65%, the recall was 91.66%, and the F-score was 91.65%. The authors' conclusion is that Support Vector Machines (SVM) yielded excellent outcomes for languages with limited resources, such as Hindi and Bengali. Another instance of a model that utilized Support Vector Machines (SVM) is presented in reference [17]. The authors recognized the following features in their papers: Title case refers to the practice of capitalizing the first letter of each word**.**

B- K-Nearest Neighbor(K-NN)

K-NN is a type of algorithm used in Supervised Machine Learning. This approach is commonly employed for regression and classification issues because to its simplicity. Being non-parametric, it does not make any assumptions about the underlying data. The K-NN technique, known for its simplicity and practicality, has been extensively employed in the fields of data mining and machine learning due to its exceptional performance. Following the training of sample data, classification is employed to predict the labels of test data points. Despite the numerous classification techniques proposed by scholars in recent decades, K-NN continues to

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

be widely utilized [18]. Research has been conducted utilizing the K-NN algorithm to do sentiment analysis on Twitter, employing both SVM and K-NN. The findings revealed that K-NN outperformed SVM in terms of results [19]. KNN-NER may get comparable results to the vanilla NER model with only 40% of the training data [20].

C- Naïve Bayes (NB)

One example of the application of Naive Bayes is a study conducted by [21], in which the Naïve Bayes Classifier was utilized to develop a Name Index translation of Hadith in the Indonesian language. The IOB Tag was applied to the dataset with a precision of 28.52%, a recall of 33.5%, and a Fscore of 31.5%. Titlecase achieved a score of 47.84% by utilizing morphological features. Regarding POS Tag and Unigram, two lexical characteristics, POS Tag demonstrated superior performance with a 76.75% result, while unigram achieved a 71.41% F1 score. The study utilized a combination of Unigram, POS Tag, and Title Case techniques, resulting in an F1 score of 82.63%. These results demonstrated a positive correlation between the utilization of more features and improved performance [22]. Conducted a study utilizing a naïve Bayes classifier to identify counterfeit news. The authors developed a software system utilizing this approach and conducted a trial on a collection of news posts sourced from Facebook. The technology achieved a 74% accuracy in correctly identifying bogus news in the test set. The research also elucidated some methods to enhance the precision. Their research shown that artificial intelligence can effectively address the issue of identifying counterfeit news.

4.2.2 Unsupervised Machine Learning

Unsupervised Named Entity Recognition (NER) methods usually depend on manually created rules and pre-existing lexicons. In Ref. [23] identify entities based on syntactic pattern matches. The Ref.[24] present an unsupervised method that is utilized in the medical field. The method initially acquires initial entities from an external source, then detects entities from sentences by employing chunking and utilizing inverse document frequency. In a similar manner, The Ref.[25] suggest utilizing knowledge bases to provide remote labels, which can then be employed to enhance the training of supervised Named Entity Recognition (NER) models. The Ref.[26] utilize a knowledge base to train a Named Entity Recognition (NER) model called KALM. This model distinguishes whether a term in a sentence is derived from the knowledge base or a conventional dictionary. Although other approaches heavily depend on external knowledge bases, Cycle NER tries to utilize a small number of entity samples without requiring external resources. A neural Hidden Markov Model (Neural HMM) is introduced in Ref. [27] for the purpose of token annotation. The Baum-Welch algorithm is used to estimate the probability distribution of the latent classes. In order to accomplish this, it depends on a collection of lexical, morphological, and syntactic characteristics that are obtained through the utilization of neural networks. CNN is utilized to extract morphological elements, whereas LSTMs are employed to capture the context of the sentence. This approach has been demonstrated to be efficacious for POS tagging. Both part-of-speech (POS) tagging and named entity recognition (NER) can be

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

framed as sequence labeling tasks. Hence, we see this method as an unsupervised reference point.

5. Evaluation Metrics for Named Entity Recognition Models

The general accuracy, or the effectiveness, of any information extraction (IE) model (including NE models) may be reliably measured by using a common formula called precision. The performance, or efficiency, may be tested by using the formula for recall. The F-measure, which is the balance or tradeoff between the precision and recall, may be measured using the F-measure formula. For evaluating a model, a set of input documents are needed, a particular set of entities to be extracted, and some value to show whether the information extracted is relevant or not [28].

The precision value is generally the portion of extracted entities that are related to the search over all of the extracted entities. The formula for precision is shown as the following:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (1)$$

The recall value is generally the portion of the extracted entities that are relevant over all of the relevant entities that exist. The formula for the recall is shown as the following:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{Fn}} \qquad (2)$$

Where Positive (P) - positive object marked as positive. Negative (N) - negative object marked as negative. False positive (FP)- negative object marked as positive. False negative (FN) - positive object marked as negative.

Finally, the F-measure shows the balance between the recall and precision. It is referred to as the harmonic mean of both of the above two formulas, which are precision and recall. The formula for F-measure may be shown as the following:

$$\text{F}_{\text{Measure}} = \frac{2\text{P} \times \text{R}}{\text{P} + \text{R}} \qquad (3)$$

It is important for a model to achieve high results after being evaluated in terms of precision and recall. If a model has a high precision and recall, it would be reliable to users, and if it has a lower precision and recall, more research needs to be carried out on the model to attempt to increase its precision and recall during evaluation.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

6. Named Entity Recognition Models in Crime Domain

This section provides a review of the current named entity recognition (NE) models that extract crime related information in the crime domain that crime analysts can use to get information that can help with analyzing crime cases. Each of the models uses different techniques. These techniques, and the information that each model extracts, are taken into consideration. The evaluation metrics of the models are also reviewed and the strengths and weaknesses of each model are given.

In the crime domain, crime analysts require information as fast as possible regarding certain crime cases so that they can try to solve the crime or prevent it from happening. Since time is very important when working with crime cases, analysts have to get information on a crime case as early as possible. They use NE models that extract crime information associated to a certain case, so that they may use this information for solving the case. Crime news documents include details on crimes, and this makes them beneficial. However, these documents are unstructured, and the beneficial information may only be extracted with use of NE models, since manually extracting this beneficial information in unstructured text is both time consuming and difficult. The NE models extract these features quicker and with high and reliable accuracy [29].

For example, In the United States, there is a NE model called Coplink that is commonly used at police stations all over the country. It is a knowledge management system where any police officer or analyst may quickly search for relationships between the features extracted from a particular crime case and other crime cases that occurred in the past. Figure 2 below presents the interface of Coplink in use by a crime analyst to search for information on a known suspect from the database.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

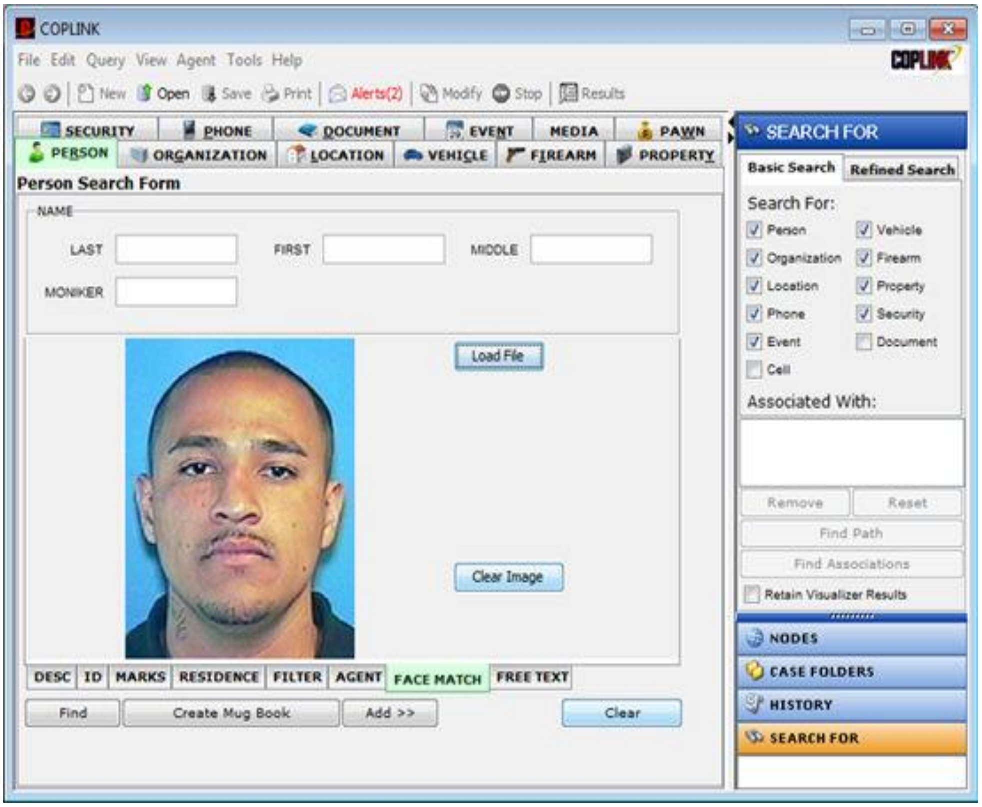
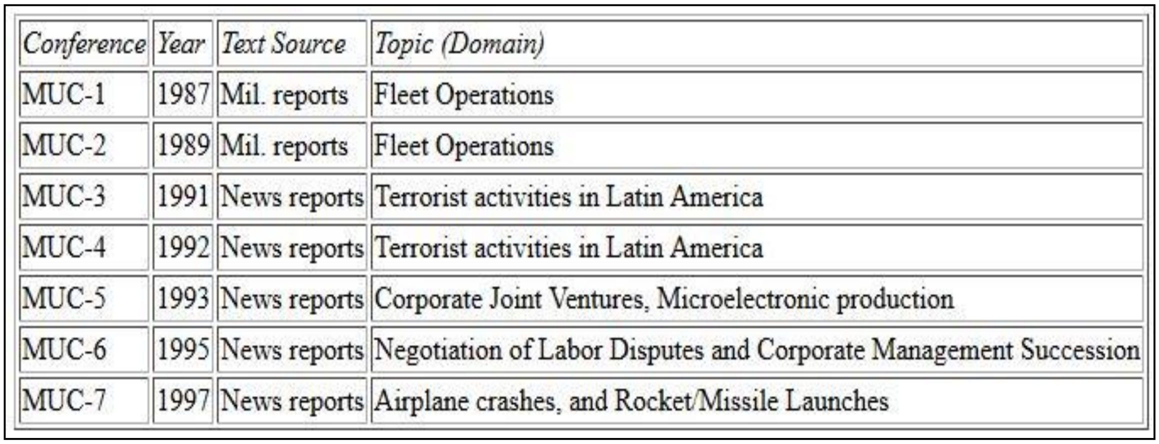

COPLINK
File Edit Query View Agent Tools Help
New Open Save Print Alerts(2) Modify Stop Results
SECURITY PHONE DOCUMENT EVENT MEDIA PAWN
PERSON ORGANIZATION LOCATION VEHICLE FIREARM PROPERTY
Person Search Form
NAME
LAST FIRST MIDDLE
MONIKER
Load File
Clear Image
DESC ID MARKS RESIDENCE FILTER AGENT FACE MATCH FREE TEXT
Find Create Mug Book Add >> Clear
SEARCH FOR
Basic Search Refined Search
Search For:
☑ Person ☑ Vehicle
☑ Organization ☑ Firearm
☑ Location ☑ Property
☑ Phone ☑ Security
☑ Event □ Document
□ Cell
Associated With:
Remove Reset
Find Path
Find Associations
□ Retain Visualizer Results
NODES
CASE FOLDERS
HISTORY
SEARCH FOR

Figure 2. Coplink User Interface

The message understanding conference (MUC) is one of the popular conferences for IE models that focus on the crime domain. It is common, and a number researcher that are examining NE models to extract crime related information are participants of this conference. It has been initialized by the Defense Advanced Research Projects Agency (DARPA) to motivate researchers to move forward in this research field so that they may propose reliable NE models that may be used by crime analysts. The NE models shown in the MUC are used for different domains, even though the crime domain is the main focus of this conference. Table 1 shows the past conferences that have taken place by the MUC [30].

Table 1. MUC conferences



Conference	Year	Text Source	Topic (Domain)
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

After the MUC, the second largest conference for IE is the Automated Content Extraction (ACE) program. The research area of IE has been now well understood by many researchers involved in the MUC, and the ACE was initialized to progress IE forward and to develop a strong foundation for IE. The ACE, at its start, had dealt with both named entity recognition and co-reference resolution, and then has moved forward to other subsections in the IE field [30]. Table 2 shows all of the different tasks that were occurred by the ACE, and the languages they were in, from the years of 2000 to 2004.

Table 2. ACE tasks

Year	Tasks	Languages
2000	EDT pilot study	English
2001	EDT, RDC	English
2002	EDT, RDC	English
2003	EDT, RDC	English, Chinese, Arabic
2004	EDR, RDR	English, Chinese, Arabic

Table 3 shows some of the existing IE and NE models that have been created for the crime domain for the years of 2002 to 2014. Each model is unique in the techniques that it uses and the entities it extracts. Some of the models rely on machine learning and rule based techniques, among others. The researcher name is shown, along with the year, techniques used, entities extracted, and the evaluation metrics for every model reviewed.

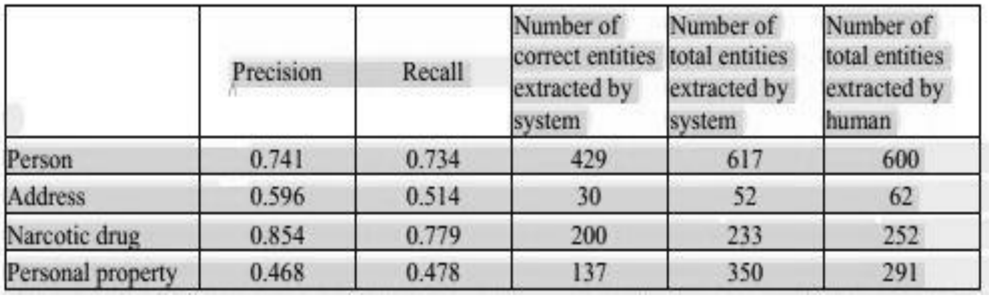
Table 3. Existing IE and NE models for crime domain

Num	Researcher	Year	Techniques Used	Entities Extracted	Evaluation Metrics
1	Chao et al.	2002	Neural networks, lexical, rule based	address, vehicle, drug, name, and property	person: P: 74%, R: 73%, drugs: P: 85%, R: 77%
2	Bengston et al.	2008	pair wise classification	nouns, noun phrases	P: 84.81%, R: 72.53%, F: 78.24%
3	Arulanandam et al.	2014	machine learning, conditional random field	Location	P: 84-90%
4	Pinheiro et al.		Semantic Inferential Model	crime type, scene	type: P: 87%, R: 71%, F: 78% scene: P: 72%, R: 68%, F: 70%
5	Alruily et al.	2009	lexical lookup, rule based	crime type	P: 60%, R: 97%, F: 74%
6	Hao et al.	2008	lexical lookup, rule based	-	P: 96%, R: 83%

In Ref. [8] had proposed a system that incorporates neural networks to obtain relative information from crime documents, reports and other text. The information that is retrieved is input to a database as a next stage for other data and text mining applications to search for patterns associated to crimes. The extracted information is in structured form, and is required for the data mining applications.

The model first incorporates neural networks, but also uses the rule based and lexical techniques. In the model, one component makes use of the Arizona Noun Phraser, and deals with noun phrases. The model can extract five named entities, which are the address, vehicle, drug, name, and property. The lexicons were generated and created for helping to extract these entities. They were gathered from the local police department in Tucson. Regarding evaluation, the model had achieved a high precision value for narcotic drugs and personal names only. The precision value for the person name and drugs was 74% and 85%. The recall values of the person name and drugs was 73% and 77%. Table 4 presents the evaluation metrics, precision and recall for the model.

Table 4. Evaluation metrics [8]



	Precision	Recall	Number of correct entities extracted by system	Number of total entities extracted by system	Number of total entities extracted by human
Person	0.741	0.734	429	617	600
Address	0.596	0.514	30	52	62
Narcotic drug	0.854	0.779	200	233	252
Personal property	0.468	0.478	137	350	291

In Ref [31] had proposed an NE model that incorporates a pair wise classification model, and uses approaches both co-reference resolution and named entity recognition. The model uses a neural network perceptron algorithm for the training of the model. The model was developed using Java programming that is based on machine learning techniques. The model was accurate with nouns and noun phrases. It searches for whether the noun phrase have any similarities with other terms following them. The model had used ACE 2004 program data set. The researchers included 336 text documents for the model evaluations. When the testing was carried out, the model had achieved a recall of 72.53%, a precision of 84.91% and an F-Measure of 78.24%.

In Ref. [32] has proposed an NE model that aims to extract information for the theft crime, and extracts the crime location (address). The theft information is obtained from newspaper articles for three countries, which are New Zealand, Australia and India. The model uses named entity recognition to show whether the sentence contains a crime location or does not. It focuses on extracting crime location from the newspaper articles. The approach used is the conditional random field, which is machine learning method to check whether or not a sentence shows crime location or not. The figure 3 below shows all of the phases that have taken place in the methodology of the system.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

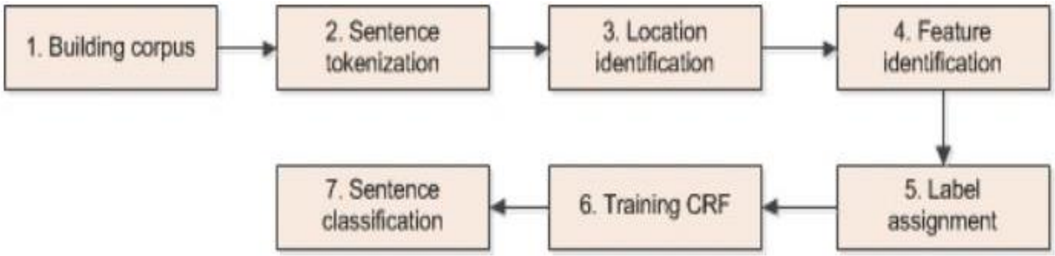
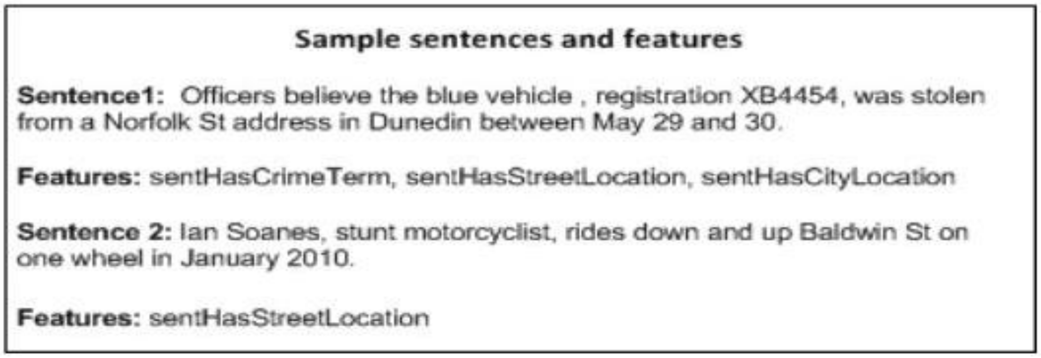

1. Building corpus
2. Sentence tokenization
3. Location identification
4. Feature identification
5. Label assignment
6. Training CRF
7. Sentence classification

Figure 3. Methodology of system

The figure 4 below presents some sample sentences and the features that are obtained for every sentence.



Sample sentences and features

Sentence1: Officers believe the blue vehicle , registration XB4454, was stolen from a Norfolk St address in Dunedin between May 29 and 30.

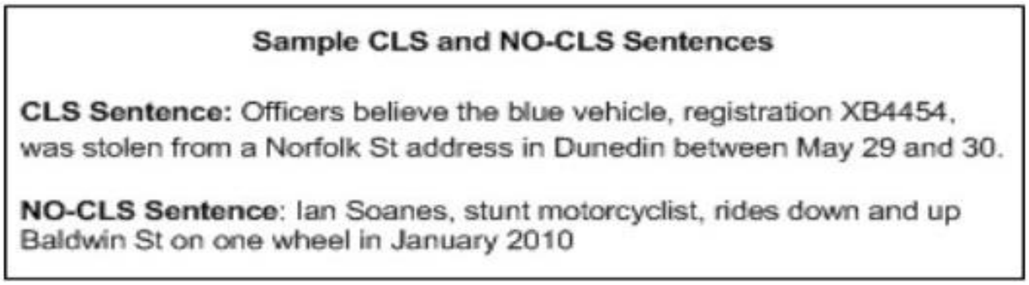
Features: sentHasCrimeTerm, sentHasStreetLocation, sentHasCityLocation

Sentence 2: Ian Soanes, stunt motorcyclist, rides down and up Baldwin St on one wheel in January 2010.

Features: sentHasStreetLocation

Figure 4. Sample sentences.

The figure 5 below shows two different sentences. One is a Crime Location Sentence, CLS, and the other is not.



Sample CLS and NO-CLS Sentences

CLS Sentence: Officers believe the blue vehicle, registration XB4454, was stolen from a Norfolk St address in Dunedin between May 29 and 30.

NO-CLS Sentence: Ian Soanes, stunt motorcyclist, rides down and up Baldwin St on one wheel in January 2010

Figure 5. Sample CLS and NO-CLS sentences.

Taking a look at the evaluation metrics, the model obtained results between 84% to 90% for New Zealand based articles, and 73% to 75% for Indian and Australian based articles.

In Ref [33] have proposed a model that relies on natural language processing methods and uses the Semantic Inferential Model. The model created is for the use of collaborative environments on the internet. The system is called Wiki Crimes, and was used to extract two main crime entities, which are crime scene and crime type from online web sites. The figure 6 below represents the architectural framework of their model.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

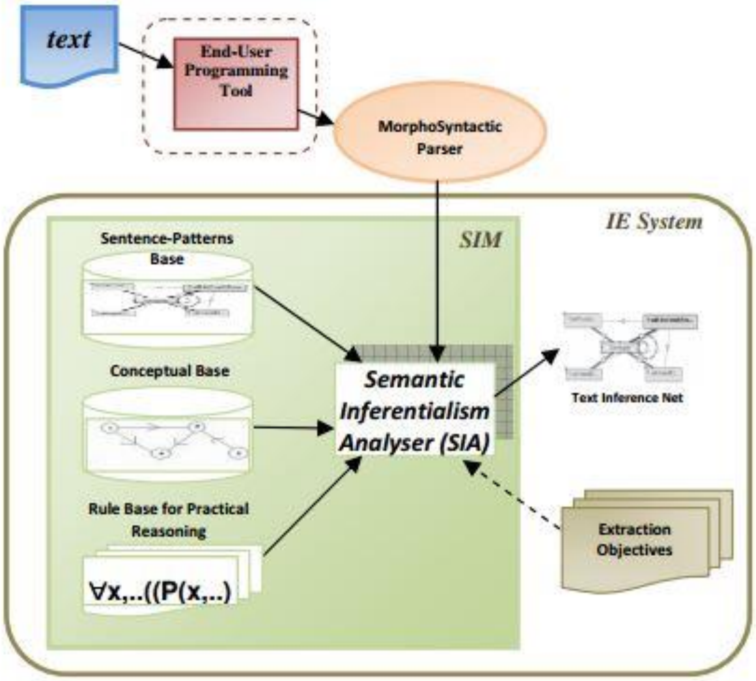
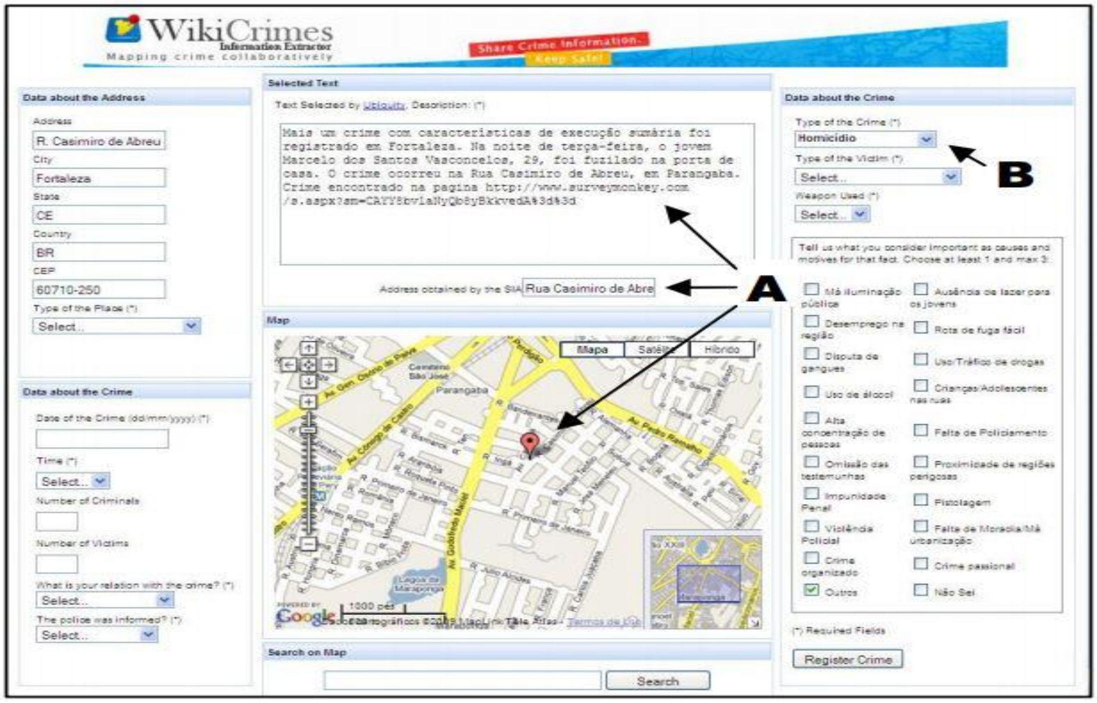


Figure 6. Model architecture.

The figure 7 presents the user interface of the system. It can be shown that the processed text was related to a crime case, and the crime scene and type were extracted from the text, and the crime scene is shown in the map provided by the user interface.



Data about the Address

Address: R. Casimiro de Abreu

City: Fortaleza

State: CE

Country: BR

CEP: 60710-250

Type of the Place (*): Select..

Data about the Crime

Date of the Crime (dd/mm/yyyy) (*):

Time (*): Select..

Number of Criminals:

Number of Victims:

What is your relation with the crime? (*): Select..

The police was informed? (*): Select..

Selected Text

Text Selected by Ubiquity. Description: (*)

```
Mais um crime com caracteristicas de execução sumária foi
registrado em Fortaleza. Na noite de terça-feira, o jovem
Marcelo dos Santos Vasconcelos, 29, foi fuzilado na porta de
casa. O crime ocorreu na Rua Casimiro de Abreu, em Parangaba.
Crime encontrado na pagina http://www.surveymonkey.com
/s.aspx?sm=CAYY8bv1aNyQb8yBkkvedA%3d%3d
```

Address obtained by the SIA: Rua Casimiro de Abre

Map

Search on Map

Search

Data about the Crime

Type of the Crime (*): Homicidio

Type of the Victim (*): Select..

Weapon Used (*): Select..

Tell us what you consider important as causes and motives for that fact. Choose at least 1 and max 3:

□ Má iluminação pública
□ Desemprego na região
□ Disputa de gangues
□ Uso de álcool
□ Alta concentração de pessoas
□ Omissão das testemunhas
□ Impunidade Penal
□ Violência Policial
□ Crime organizado
☑ Outros

□ Ausência de lazer para os jovens
□ Rota de fuga fácil
□ Uso/Tráfico de drogas
□ Crianças/Adolescentes nas ruas
□ Falta de Policiamento
□ Proximidade de regiões perigosas
□ Pistolagem
□ Falta de Moradia/Má urbanização
□ Crime passional
□ Não Sei

(*) Required Fields

Register Crime

Figure 7. Wiki Crimes Interface.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Regarding the evaluation metrics, the model had obtained a precision, recall and F-measure of 87%, 71% and 78% for crime scene. It had achieved a precision, recall and F-measure of 72%, 68% and 70% for type of crime.

In Ref. [34] created a system entitled the Crime Type Recognition System (CTRS) that obtains crime type and makes use of two techniques, which are the lexical lookup technique and the rule-based technique. The first technique is the use of gazetteers and lexicons of verbs and names. The second technique is rule-based and relies on several rules and the crime indicator list to successfully obtain the type of crime. The proposed model is built to deals with Arabic texts and produced satisfying results, where the precision (60%), recall (97%), and F-measure values (74%).

The primary stage of the model is to conduct the preprocessing, and this consists of the removal of punctuation, comma, and stop words. After this, tokenization occurs. Later, the model searches the database that includes list of crime verbs and list of crime names. It attempts to relate any words in these lists with the text document that was input. The CTRS first attempts to use the lexicons and gazetteers to search for crime related information, and if nothing was found, it would use the rule based method. The model works in languages of both the English and Arabic. They had also made use of the lexicons in both languages. News documents were used for testing the model. After the evaluation of the model, it had achieved the precision value of 60%, recall value of 97% and F-score of 74%. Table 5 below shows the evaluation metrics of the model.

Table 5. Evaluation metrics of model [34]

Evaluation type	**Result**
Precision	60%
Recall	97%
F-measure	74%

In Ref. [35] also created another similar system that recognized other crime related information other than the type of crime. The newer system has the capability to extract the nationality of the crime, the location, and similarly to the previous model, the crime type. For the Arabic language, the system makes use of an indicator, which did not work correctly for English because in English there are multiple indicator terms that may be implemented which may come after the term or before it.

In Ref. [36] had proposed a model that retrieves crime related information from unstructured text. Their IE model uses many natural language processing techniques to find

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652

useful information from news articles and witness and police reports. Their model makes use of the methods of rule based and lexical lookup. The lexicons were extremely large and were generated by checking several different areas such as the Uniform Crime Reports, FBI, Frame Net, Wikipedia, among other sources. For the model creation, the General Architecture for Text Engineering (GATE) was implemented. Figure 8 below presents the framework of the model.

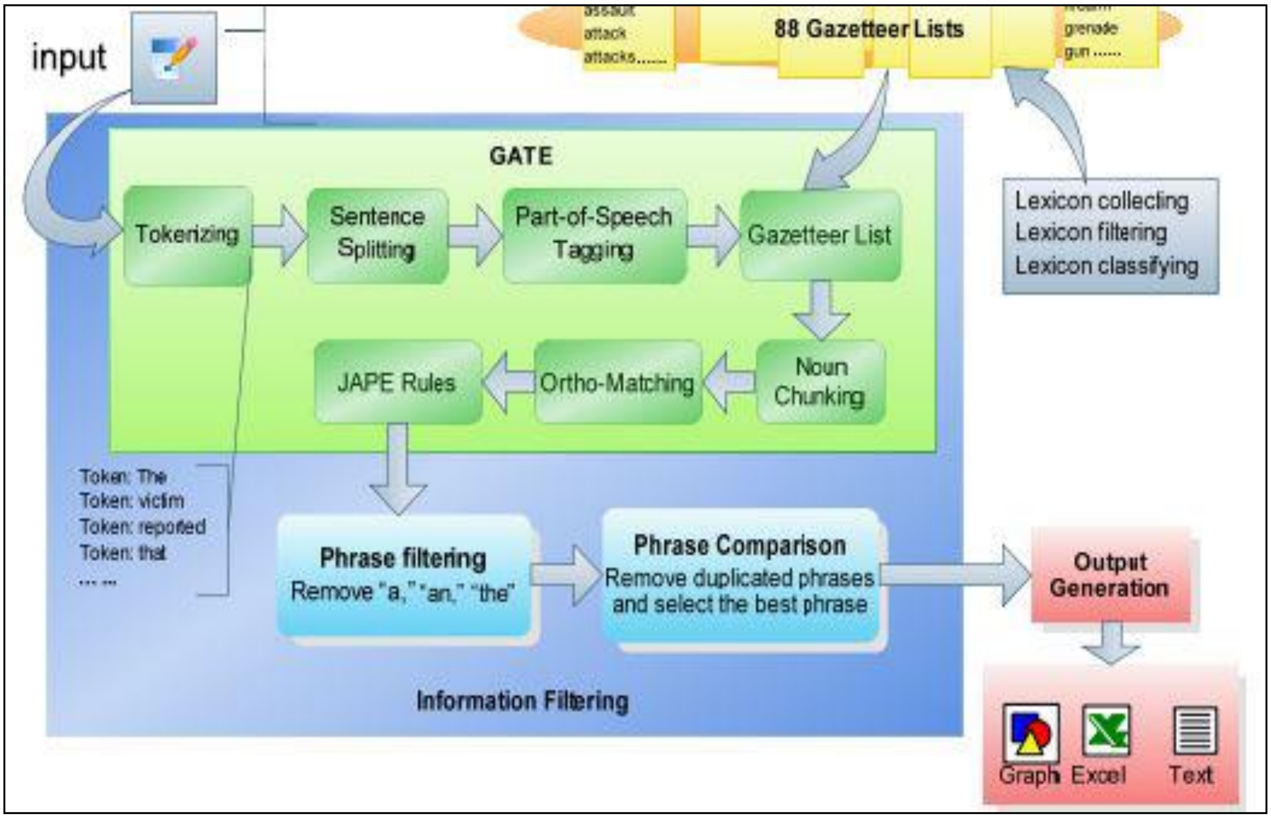


Figure 8. Framework of Model [36]

After the model evaluation and testing, it had obtained a precision of 96% and a recall of 83% during the processing of the police reports. The model also was evaluated the witness reports, where it obtained less precision and recall value, which were 93% for precision and 77% for recall.

After a review of the related work on NE models, it can be concluded that there still exists a gap, since an NE model that extracts crime related information at the level of human performance and accuracy has not yet been developed. Further research on the development of an NE model that extracts crime related information that is beneficial to crime analysis, and performs at the level of humans in terms of effectiveness and efficiency is needed to be carried out to fill this gap.

مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية مجلة جامعة بابل للعلوم الصرفة والتطبيقية

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

7. Conclusion

Named Entity Recognition is a technique of natural language processing that is used for the categorization of the data. It is a subfield of Artificial Intelligence(AI) and is being used heavily in the industries today to automate major categorization of data for unstructured text and datasets. Understanding NER in reference to NLP (natural language processing) is pretty simple. It is one of the key detection techniques used in NLP. The capability of a system to identify key elements or entities such as name, organization, designation, time, experience, etc. comes under the bracket of NER. Without NER, data extraction using NLP will be devoid of any context that is required to understand a phrase, a sentence, or an article to effectively process that data. With NER, it is possible to search key elements or entities from a phrase, sentence, or a paragraph. This capability can be used during online searches to find elements such as name, organization, events, locations, etc. from thousands of articles which otherwise would require tons of resources at hand.

Conflict of interests.

There are non-conflicts of interest.

References

[1] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, *et al.*, "Unified structure generation for universal information extraction," *arXiv preprint arXiv:2203.12277*, 2022.

[2] H. Reddy, N. Raj, M. Gala, and A. Basava, "Text-mining-based fake news detection using ensemble methods," *International Journal of Automation and Computing*, vol. 17, pp. 210-221, 2020.

[3] R. Grishman, "Twenty-five years of information extraction," *Natural Language Engineering*, vol. 25, pp. 677-692, 2019.

[4] H. Shabat, N. Omar, and K. Rahem, "Named entity recognition in crime using machine learning approach," in *Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings 10*, 2014, pp. 280-288.

[5] A. S. Sidhu, "Crime levels and trends in the next decade," *Journal of the Kuala Lumpur Royal Malaysia Police College*, vol. 5, pp. 1-13, 2006.

[6] K. R. Rahem and N. Omar, "RULE-BASED NAMED ENTITY RECOGNITION FOR DRUG-RELATED CRIME NEWS DOCUMENTS," *Journal of Theoretical & Applied Information Technology*, vol. 77, 2015.

[7] H. Cunningham and S. Azzam, "Information Extraction," *Automatic, Encyclopedia of Language and Linguistics*, 1995.

[8] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," 2002.

[9] M.-C. De Marneffe and C. D. Manning, "Stanford typed dependencies manual," Technical report, Stanford University2008.

[10] L. Xiao, D. Wissmann, M. Brown, and S. Jablonski, "Information extraction from the web: System and techniques," *Applied Intelligence,* vol. 21, pp. 195-224, 2004.

[11] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: a systematic review," *Computer Science Review,* vol. 29, pp. 21-43, 2018.

[12] O. Ramos Flores and D. Pinto, "Proposal for named entities recognition and classification (NERC) and the automatic generation of rules on Mexican news," *Computación y Sistemas,* vol. 24, pp. 533-538, 2020.

[13] R. Alfred, L. C. Leong, C. K. On, P. Anthony, T. S. Fun, M. N. B. Razali, *et al.*, "A rule-based named-entity recognition for malay articles," in *Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part I 9*, 2013, pp. 288-299.

[14] A. Anandika and S. P. Mishra, "A study on machine learning approaches for named entity recognition," in *2019 International Conference on Applied Machine Learning (ICAML)*, 2019, pp. 153-159.

[15] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review,* vol. 47, pp. 279-311, 2017.

[16] A. Ekbal, S. Saha, and D. Singh, "Active machine learning technique for named entity recognition," in *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 2012, pp. 180-186.

[17] F. A. Yusup, M. A. Bijaksana, and A. F. Huda, "Narrator's name recognition with support vector machine for indexing Indonesian hadith translations," *Procedia Computer Science,* vol. 157, pp. 191-198, 2019.

[18] A. Pandey and A. Jain, "Comparative analysis of KNN algorithm using various normalization techniques," *International Journal of Computer Network and Information Security,* vol. 11, p. 36, 2017.

[19] M. R. Huq, A. Ahmad, and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *International Journal of Advanced Computer Science and Applications,* vol. 8, 2017.

[20] S. Wang, X. Li, Y. Meng, T. Zhang, R. Ouyang, J. Li, *et al.*, "$ k $ NN-NER: Named Entity Recognition with Nearest Neighbor Search," *arXiv preprint arXiv:2203.17103,* 2022.

[21] F. Y. Azalia, M. A. Bijaksana, and A. F. Huda, "Name indexing in Indonesian translation of hadith using named entity recognition with naïve bayes classifier," *Procedia Computer Science,* vol. 157, pp. 142-149, 2019.

[22] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, 2017, pp. 900-903.

[23] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, *et al.*, "Unsupervised named-entity extraction from the web: An experimental study," *Artificial intelligence,* vol. 165, pp. 91-134, 2005.

[24] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics,* vol. 46, pp. 1088-1098, 2013.

[25] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9143-9150.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

[26] A. Liu, J. Du, and V. Stoyanov, "Knowledge-augmented language model and its application to unsupervised named-entity recognition," *arXiv preprint arXiv:1904.04458*, 2019.

[27] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[28] F. M. Kamau, K. Ogada, and C. W. Kipruto, "Analysis of Machine-Based Learning Algorithm Used in Named Entity Recognition," *Informing Science: The International Journal of an Emerging Transdiscipline*, vol. 26, pp. 069-084, 2023.

[29] N. Kumar and P. Bhattacharyya, "Named entity recognition in Hindi using MEMM (Technical Report)," *IIT Mumbai*, 2006.

[30] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

[31] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 294-303.

[32] R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis, "Extracting crime information from online newspaper articles," in *Proceedings of the second australasian web conference-volume 155*, 2014, pp. 31-38.

[33] V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text," in *2010 IEEE International Conference on Intelligence and Security Informatics*, 2010, pp. 19-24.

[34] M. Alruily, A. Ayesh, and H. Zedan, "Crime type document classification from arabic corpus," in *2009 Second International Conference on Developments in eSystems Engineering*, 2009, pp. 153-159.

[35] M. Alruily, A. Ayesh, and A. Al-Marghilani, "Using self organizing map to cluster arabic crime documents," in *Proceedings of the international multiconference on computer science and information technology*, 2010, pp. 357-363.

[36] C. H. Ku, A. Iriberri, and G. Leroy, "Crime information extraction from police and witness narrative reports," in *2008 IEEE Conference on Technologies for Homeland Security*, 2008, pp. 193-198.