Journal of Theoretical and Applied Information Technology
20th July 2015. Vol.77. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RULE-BASED NAMED ENTITY RECOGNITION FOR DRUG-RELATED CRIME NEWS DOCUMENTS

[1]KHMAEL RAKM RAHEM, [2]NAZLIA OMAR
Center for AI Technology (CAIT), Faculty of Information
Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
E-mail: [1]khm2006_rakem@yahoo.com , [2]nazlia@ukm.edu.my

ABSTRACT

Drug abuse pertains to the consumption of a substance that may induce adverse effects to a person. In international security studies, drug trafficking has become an important topic. In this regard, drug-related crimes are identified as an extremely significant challenge faced by any community. Several techniques for investigations in the crime domain have been implemented by many researchers. However, most of these researchers focus on extracting general crime entities. The number of studies that focus on the drug crime domain is relatively limited. This paper mainly aims to propose a rule-based named entity recognition model for drug-related crime news documents. In this work, a set of heuristic and grammatical rules is used to extract named entities, such as types of drugs, amount of drugs, price of drugs, drug hiding methods, and the nationality of the suspect. A set of grammatical and heuristic rules is established based on part-of-speech information, developed gazetteers, and indicator word lists. The combined approach of heuristic and grammatical rules achieves a good performance with an overall precision of 86%, a recall of 87%, and an F1-measure of 87%. Results indicate that the ensemble of both heuristic and grammatical rules improves the extraction effectiveness in terms of macro-F1 for all entities.

Keywords: *Named Entity Recognition, Rule-Based Approach, Grammatical Rules, Heuristic Rules*

1. INTRODUCTION

Named entity recognition (NER) extracts valuable entities from unstructured data, such as the names of people, location, organization, and others (e.g., date, time, percentage, and monetary expressions). These data can be extracted from documents even without a “deep understanding” of the text.

Valuable crime information can be obtained from human-readable text format in online newspapers and electronic archives. Nevertheless, the number of software systems that can extract and provide relevant information remains scarce. This particular detail has significantly received wide attention from researchers in the field of information extraction. The number of drug users in Malaysia is increasing [1], and narcotics abuse by individuals is correlated with crime in general [2]. Thus, drug trafficking has become an important topic in international security studies.

Named entity extractions have different approaches, such as rule-based approach, machine learning, and hybrid approach. All of these methods attempt to recognize the necessary named entities. Rule-based model depends on human-made rules to recognize the required named entities. The application of rule-based models is far from a simple task; skilled linguists must deal with the model to give it suitable corrections [3]. This technique obtains high accuracy when used with certain domains [4]. Rule-based approaches are domain dependent, making them appropriate for NER in the crime domain.

Named entity recognition systems for general entities can extract only entities that are common in many domains; hence, specific entities in a particular task must be separately extracted because of their importance. However, the number of studies on drug crime domain is relatively limited in extracting information from documentation of drug crimes. Analyzing and extracting such drug crime information from online articles pose numerous issues. First, extracting uncommon specific entities, which are usually related to only a definite domain, is an exceedingly challenging task for generic extraction models.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Second, several online data are unstructured and overly noisy; they contain various spelling and typographical errors and grammatical mistakes, making the extraction of entity more difficult. Third, different from extracting common entities (e.g., names of person and locations), deriving drug-related crime entities (e.g., drugs hiding methods) is a complicated task because they are not written in standard and specific patterns.

We believe that a research on crime domains that specifically focuses on drug crimes would be a success. This work is mainly motivated by the need to develop a successful and useful NER system for crimes. Developing a tool that helps the community and enforcement agencies to extract drug-related crime information is significantly needed. This tool should filter the vast amount of knowledge available in the web, use indexed reliable sources such as online newspapers and blogs, and extract drug-related information. Thus, this study primarily aims to extract and analyze available drug-crime information in online newspaper articles. This work develops the following: (1) several drug-dependent specific linguistic resources such as drug-crime-specific lexicons and gazetteers; and (2) a rule-based extraction tool to assess where and how drug traffickers hide drugs, identify the nationalities of drug suspect and the types (names) of drugs, and determine the quantity and prices of drugs in the local market.

2. RELATED WORK

This section highlights recent works that extract crime information from unstructured documents. [5] proposed named-entity extraction model that uses neural networks as well as rule-based and lexical techniques. The model can recognize a total of five entities, which are the address, vehicle, name, narcotic drug, and property. The precision value for the name, address, property, and narcotic drug was 74%, 59%, 49%, and 85%, respectively.

[6] created a model that extracts crime-related information from witness and police reports. Their model makes use of the approaches of rule-based and lexical lookup. The lexicons were very large; they were gathered by checking many different sources, such as the FBI, Uniform Crime Reports, Frame Net, and Wikipedia. The extracting entities are people, act, scene, time, age, face, body part, personal property, physical condition, vehicle, clothes, weapon, feature, drug, and hair. For the model construction, the general architecture for text engineering (GATE) was used. The model achieved a precision of 96% and a recall of 83% during the processing of the police reports. In addition, the model was tested on the witness reports, where it obtained less precision and recall values, which were 93% and 77%, respectively.

[7] proposed a model that extracts crime type, location, and nationality for Arabic. This model involves adaptive dictionary building. The results of crime type are 90.25% precision, 50.47% recall, and 64.74% F-measure. The results of location are 70.91% precision, 50.47% recall, and 64.74% F-measure. The results of nationality are 81.94% precision, 95.93 % recall, and 88.38 % F-measure.

[8] extracted named entities from Arabic crime documents giving general information for crime analysis. This model utilizes a rule-based technique. On the basis of morphological information, predefined crime indicator lists and a named entity dataset in the crime domain, these rules recognize and classify entities, such as names, locations, organizations, times and dates in crime documents. The overall results of their model are 89% precision, 88% recall, and 90% F-measure.

[9] proposed an evaluation of both direct and indirect identification of the nationality of victims or suspects from online crime news documents. Direct extraction extracts the explicitly mentioned nationality within the crime documents. If the nationality cannot be extracted with use of direct extraction, then indirect extraction identifies any indirectly mentioned nationalities. It requires to access to lexicon lists and lists of indicator keywords. The results of direct with indirect extraction are 55% precision, 96% recall, and 70% F-measure. The results of victim–suspect reference identification are 62% precision, 53% recall, and 57% F-measure.

[10] introduced a relatively newer work where an information extraction model focuses on the extraction of information for one type of crime, which is theft, and extracts crime location. Theft-related information is extracted from newspaper articles from three different countries, which are New Zealand, Australia, and India. The model uses NER to show if the sentence contains a crime location or not.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The literature shows that the common entities extracted in crime domains include the names of victims, weapons, and drugs, which can be used to achieve effective results. This work proposes a new model that uses rule-based approaches in drug crimes to extract drug-related crime information.

3. ARCHITECTURE OF THE SYSTEM

This section provides the general methodology for extracting and analyzing drug-related crime information from newspaper articles. Several techniques are adopted in each phase to extract related entities. Figure 1 shows the architecture of the system.

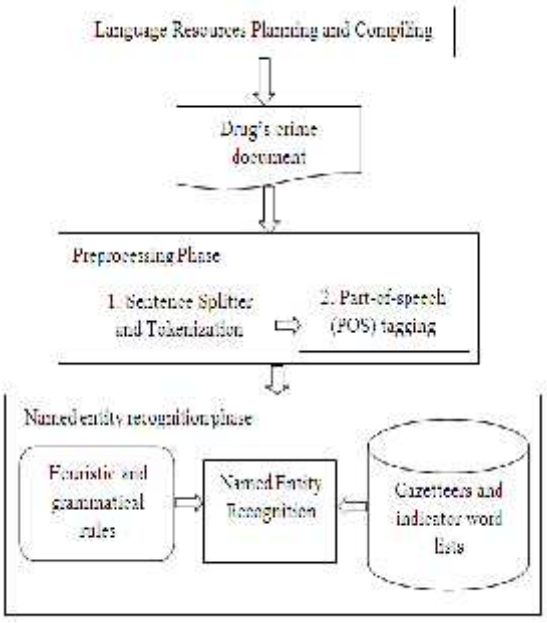

Figure 1: Proposed architecture and components.

4. LANGUAGE RESOURCES PLANNING AND COMPILING

All data used in this study are related to the drug-crime domain. The data source was collected from the national news agency of Malaysia (BERNAMA). The documents gathered were in TXT format. After the collection of the data set, the information gathered was analyzed, and a set of resources was developed. The resources developed are discussed as follows:

Indicator Words Lists: The indicators form a window around a named entity, which helps identify named entities within the text but does not get recognized itself [11]. Introductory lists play a central role in the development of the heuristics [12]. The developing indicator word lists are price indicator word list (PIW), quantity indicator word list (QIW), nationality indicator word list (NIW), and drug-hiding methods indicator word list (DHMIW).

Gazetteer: The gazetteer is a set of lists that contains specific information, such as names of people, location, organizations, and days of the week [8]. Three types of lists are used in the proposed NER rule-based model. A drug list (DL) contains the names of drugs, a nationality list (NL) contains the names of nationalities, and an extra nationality list (ENL) contains some additional nationalities not existing in the NL.

5. PRE-PROCESSING PHASE

Pre-processing is very important task in any NER system. This sub-task provides syntactic and semantic analyses of texts as inputs to the event detector modules [13]. In this work, the Stanford Natural Language Processing (NLP) tool was selected. The Stanford NLP tool is an open source library which is comprised of several modules. The modules adopted in this study include the sentence splitter and tokenisation and parts of speech (POS) tagger.

sentence splitter and tokenization: The sentence splitter split the text into several sentences [6]. The tokenizer splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case) [14].

Part of speech tagger: This task computationally determines the POS of a word and how it is activated by use in a particular context. Each token is annotated with a tag. The tags refer to the grammatical category, e.g., verb, noun, adjective. In this step, each word will receive a unique tag. Figure 2 shows a sample of POS output.

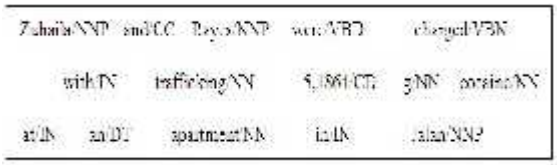

Figure 2 : A sample of POS output

Journal of Theoretical and Applied Information Technology
20th July 2015. Vol.77. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. NAMED ENTITY RECOGNITION PHASE

6.1 Heuristic Rules

In the proposed work, heuristic rules are applied using several algorithms to extract named entities.

6.1.1 drug name extraction

Algorithm 1 was applied to extract the name of drug. Each word in a crime document is compared to the drug names in the DL. If there are matches, the name of the drug will be extracted.

Algorithm 1 Drug name extraction

```
1   Require: file ≠0
2   Token <= file
3   DL <= Drug List
4   while file ≠0 do
5   For DLcounter=0 to length [DL]-1 do
6   If (token= =DL[DLcounter]) or
    (previousToken+ token= =DL[DLcounter] ) ) do
7   Drug Extracted <= token
8   End if
9   End for
10  End while
```

6.1.2 quantity of drug extraction

Algorithm 2 was applied to extract the drug quantity. The QIW list was used in the extraction process. If a numeric value referring to quantity of drug occurs first in the text followed by quantity indicator word, then the quantity of drug entity was extracted.

Algorithm 2 Quantity of drug extraction

```
1   Require: file ≠0
2   Token <= file
3   QIW <= Quantity Indicator Word List
4   while file ≠0 do
5   previousToken <= Token[-1]
6   For QIWcounter=0 to length [ QIW ]-1 do
7   If ( previousToken== number and Token==
    QIW[QIWcounter])
8   Drugs Quantity Extracted <= previousToken + token
9   End if
10  End for
11  End while
```

6.1.3 price of drug extraction

Algorithm 3 was applied to extract the price of the drug. The PIW list was used in the extraction process. If the price indicator word occurs first in the text followed by a numeric value referring to a price of drug, then the price of drug entity was extracted.

Algorithm 3 Price of drug extraction

```
1   Require: file ≠0
2   Token <= file
3   PIW <= Price Indicator words list
4   While file ≠0 do
5   previousToken(s) <= Token[-1…-3]
6   For PIWcounter=0 to length [ PIW ]-1 do
7   If ( previousToken(s)== PIW[ PIWcounter]
    and Token== number)
8   Drugs Price Extracted <= "RM" + token
9   End if
10  End for
11  End while
```

6.1.4 nationality of suspect extraction

To extract a drug-dealer's nationality, Algorithm 4 was applied. This algorithm comprises two steps. First, all nationalities in the document are collected based on the NL and ENL lists. Second, a filtration occurs to determine the nationality of suspect. This extraction is conducted by using the NIW to determine all the indicator words in the text that are drug-dealer's related. Thereafter, the distance of each nationality indicator word is calculated in relation to the position of the nationality that has been extracted. The nationality with the smallest distance is returned.

Algorithm 4 Nationality of suspect extraction

```
1   Require: file ≠0
2   Token <= file
3   NL <= Nationality List
4   ENL <= Extra Nationality Keywords List
5   NIW <= Nationality Indicator Keywords List
6   Extracted Nationalities <= empty list
7   while file ≠0 do
8   previousToken <= Token[-1]
9   For NLcounter=0 to length [NL]-1 do
```

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
10  If(token==NL[NLcounter])or(previousToken+token==
                NL[NLcounter])) do
11    Nationality Extracted <= token
12    ExtractedNationalities.add( Nationality extracted)
13   End if
14   End for
15        For ENL counter=0 to length [ENL]-1 do
16        If ( token= = ENL [ENL counter]) do
17             Nationality Extracted <= token
18             ExtractedNationalities.add( Nationality
     extracted)
19         End if
20        End for
21    End while
22     while file ≠0     do
23  For ENcounter = 0 to length[ExtractedNationalities]- 1 do
24  For NIWcounter=0 to length [NIW]-1 do
25  If(token==NIW[NIWcounter] or previousToken+token
     == NIW[NIWcounter] )
26   Distance=compute-distance (token,
     ExtractedNationalities[ENcounter] )
27   Distancelist[ENcounetr]= Distancelist[ENcounter]
     + Distance
28             End if
29          Endfor
30    Endfor
31  End while
32  Return Extracted Nationality with smallest distance
```

6.1.5 drug-hiding methods extraction

Algorithm 5 was applied to extract the drug-hiding methods. If drug hiding methods indicator word exists in the text, followed by hidden method then the drugs hiding method entity was extracted.

Algorithm 5 Drug-hiding methods extraction

```
1   Require: file ≠0
2   Token <= file
3   DHMIW <= Drugs Hiding Methods Indicator  words List
4       previousToken(s) <= Token[-1…-3]
5    while file ≠0     do
6      For DHMIW counter=0 to length [DHMIW ]-1 do
7      If (previous Token(s) == DHMIW [DHMIW counter])
8           NextTokens <= Token[1…2]
9   Drugs hiding method extracted   <= Token + Next Tokens
10           End if
11       End for
12    End while
```

6.2 Grammatical Rules

The grammatical rules are a set of grammatical patterns designed to derive drug-crime entities based on POS information. These patterns are indicated by regular expressions. Several rules have been developed, which are used to extract the quantity, price, and drugs hidden methods entities. The examples below show some of these cases.

1-**quantity rule**: <CD>{<NN>|<NNS>|(QIW)}
{<IN >}?{<NN>|<NNS|(DL)}

Examples:

- < CD > /201 QIW|< NNS >/ grammes <IN >/ of <NN >|DL/ heroin
- <CD >/105 QIW|< NN >/g <NN >|DL/cannabis
- <CD >/ 46< NN >/syabu <NNS >/ pills

1- **price rule:** <IN>{< DT >|<IN>}?{<NNP>}? < CD >*

Examples:

- <IN>/ worth <NNP>/ RM <CD>/3000
- <IN>/worth<DT>/some<NNP>/RM <CD>/100,000
- < IN >/ worth < IN >/ about < NNP >/ RM <CD>/35,000

2- **hiding-methods rule**: <NNS> <VBD> CD>? <DT>? <JJ>? <NNS>{<IN> {<NN>| NNS> |(DL)}}?

Examples:

- <CD>/ two <NNS >/ bottles<IN>/ of <NN>|DL/syabu
- <DT>/a<NN>/ packet <VBG>/containing <NNS>|DL/cannabis
- <CD>/two<JJ>/ plastic <NNS>/bags<VBG>/ containing<NN>|DL/syabu

7. EVALUATION

This work used a golden standard that was manually developed, where all entities were manually annotated. A total of 200 online drug crime news documents were gathered from the Malaysian national news agency (BERNAMA).

Journal of Theoretical and Applied Information Technology
20th July 2015. Vol.77. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Thirty random documents were selected as a test set to be included in the experiments performed on the system.

The performance measures used to evaluate the entity extraction system are precision, recall, and weighted mean F-measure. Precision is the percentage of correct entities found by the system. Recall is the percentage of entities present in the corpus that are found by the system. F-measure is the value of the harmonic mean of both the precision and the recall [6], [14]. The definitions of the performance measures used are summarized below.

$$\text{Recall} = \frac{\#\text{ of correctly classified drugs NEs}}{\#\text{ of drugs NEs in the corpus}} \quad (1)$$

$$\text{Precision} = \frac{\#\text{ of correctly classified drugs NEs}}{\#\text{ of drugs NEs in the system}} \quad (2)$$

$$\text{F- Measure} = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \quad (3)$$

8. RESULTS AND DISCUSSION

Two experiments were applied to evaluate the NER system in terms of efficiency and accuracy. The first experiment applied heuristic rules to extract the named entities. The overall results for this experiment are 85% precision, 75% recall, and 79% F-measure. The second experiment applied heuristic and grammatical. The overall results for this experiment are 86% precision, 87% recall, and 87% F-measure. The evaluation metrics for each entity in the final experiment are shown in Table 1.

Table 1: Final results of proposed system.

Entity name	Precision (%)	Recall (%)	F-measure (%)
Name of drugs	96	94	95
Price of drugs	79	94	86
Quantity of drugs	89	82	85
Nationality of suspect	92	89	90
Method of hiding drugs	76	78	77

The results are enhanced in the second experiment by adding several grammatical rules to the heuristic rules. Figure 3 shows the comparison of overall evaluation metrics for each experiment.

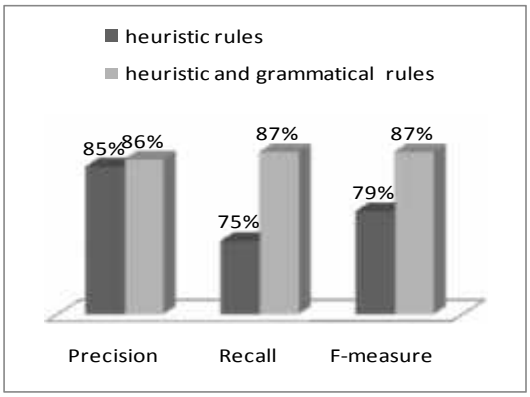

Figure 3: Comparison of evaluation metrics for the system experiments

Also, Drug names and nationality of suspect have the highest results because most drug names and nationalities are limited and can be listed. Extracting the methods of hiding drugs has the worst result because hiding methods are boundless (unlimited) and cannot be listed. Furthermore, is not written in standard and common patterns.

9. CONCLUSION

This work developed a rule-based NER system to collect relevant drug-crime information to help community and police investigators. This study revealed several drug-dependent-specific linguistic resources such as drug-crime-specific lexicon and gazetteers. Moreover, a rule-based extraction tool was established to assess where and how drug dealers hide drugs, identify the nationalities of suspects and the names of drugs, and evaluate the quantity and prices of drugs in the local market. The final experiment indicated promising results and demonstrated that a rule-based method is suitable for extracting drug-related crime entities.

REFRENCES:

[1] G. Reid, A. Kamarulzaman and S. K. Sran, "Malaysia and Harm Reduction: The Challenges and Responses", *International Journal of Drug Policy*,Vol. 18, Issue. 2, 2007, pp.136–140.

[2] A. S. Sidhu, "The Rise of Crime in Malaysia An Academic and Statistical Analysis", *Journal of the Kuala Lumpur Royal Malaysia Police College*, No. 4, 2005, 1-28.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] J. Mayfield, P. McNamee and C. Piatko, "Named Entity Recognition using Hundreds of Thousands of Features", *In: Proceedings of the 7th conference on Natural language learning at HLT-NAACL (CONLL 2003)*, vol. 4, 2003, pp.184–187.

[4] K. Shaalan, "Rule-based approach in Arabic natural language processing", *The International Journal on Information and Communication Technologies (IJICT)*, Vol. 3, No. 3, June 2010, pp. 11–19.

[5] M. Chau, J. J. Xu, and H. Chen, "Extracting Meaningful Entities From Police Narrative Reports". *In Proceedings of the 2002 annual national conference on Digital government research Anjuran Digital Government Society of North America*. Los Angeles, California, May 2002, pp.1-5.

[6] C. H Ku, A. Iriberri and G. Leroy, "information extraction from police and witness narrative reports", *Technologies for Homeland Security Conference*, Waltham, MA, May 12-13, 2008, pp.193-198.

[7] M. Alruily, A. Ayesh and H. Zedan, "Automated dictionary construction from Arabic corpus for meaningful crime information extraction and document classification". *In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*, IEEE, 2010, pp. 137-142.

[8] M. Asharef, N. Omar and M. Albared, "Arabic Named Entity Recognition in Crime Documents", *Journal of Theoretical and Applied Information Technology*, Vol. 44, No. 1, 2012, pp.1-6.

[9] A. Alkaff and M. Mohd, "Extraction of Nationality from Crime News", *Journal of Theoretical and Applied Information Technology*, Vol. 54, No. 2, 2013, pp.304-312.

[10] R. Arulanandam, B. T. R. Savarimuthu and M. A. Purvis, "Extracting Crime Information from Online Newspaper Articles", *Proceedings of the Second Australasian Web Conference (AWC 2014)*, Auckland, New Zealand, 2014, pp.31-38.

[11] K. Shaalan and H. Raza, "NERA: Named Entity Recognition for Arabic", *Journal of the American Society FOR Information Science and Technology*, vol. 60, issue. 8, 2009, pp. 1652–1663.

[12] A. Elsebai, F. Meziane and FZ. BelKredim, "A Rule Based Persons Names Arabic Extraction System", *Communications of the IBIMA*, Vol. 11, 2009, 11: 53–59.

[13] M. Amami, A., Elkhlifi and R. Faiz, "BioEv: A System for Learning Biological EventExtraction", *International Conference on Information Technology and e-Services (ICITeS)*, Sousse , 24-26 March 2012, pp.1-5.

[14] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan and H. Cunningham, "Shallow methods for named entity coreference resolution", *InChaınes de références et résolveurs d'anaphores*, workshop TALN, 2002.

[15] H.Cunningham,"Information Extraction", Automatic, *Encyclopedia of Language & Linguistics*, Second Edition, vol 5, Oxford, Elsevier, 2006. pp. 665-677.