Probabilistic Reference to Suspect or Victim in Nationality Extraction from Unstructured Crime News Documents

Mohammad Darwich Masnizah Mohd
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia

Abstract

There is valuable information in unstructured crime news documents which crime analysts must manually search for. To solve this issue, several information extraction models have been implemented, all of which are capable of being enhanced. This gap has created the motivation to propose an enhanced information extraction model that uses named entity recognition to extract the nationality from crime news documents and coreference resolution to associate the nationality to either the suspect or the victim. After the proposed model extracts the nationality, it references it to the suspect or victim by looking up all of the victim related keywords and the suspect related keywords within the text, and their corresponding distances from the position of the nationality keyword. Based on their total distances, a probability score algorithm decides whether the nationality is more likely to belong to either the victim or the suspect. Two experiments were conducted to evaluate the nationality extractor component and the reference identification component used by the model. The former experiment had achieved 90%, 94%, and 91% for precision, recall, and F-measure values respectively. The latter experiment had achieved 65%, 68%, and 66% for precision, recall, and F-measure respectively. The model had achieved promising results after evaluation.

Keywords: information extraction, named entity recognition, coreference resolution, crime domain

1. Introduction

In the crime domain, it is critical that crime analysts and investigators access criminal justice data and intelligence on crime cases efficiently (in terms of speed) and effectively (in terms of accuracy) to perform investigations and prevent crime. There is valuable information in online crime news documents, which usually contain text that is unstructured. Valuable information are entities within the text, which may be person names, nationalities, crime locations, crime dates, crime types, criminal properties, weapons used, narcotic drugs, car brand, among others (Chau et al. 2002; Feldman et al. 2006). Information Extraction (IE) systems are used for solving this issue.

2. Information Extraction Models in the Crime Domain

Several IE models (or systems) have been implemented in order to keep analysts and investigators updated with the information they need accurately and efficiently (Jurafsky et al. 2009). These systems (Chao et al. 2002; Hao et al. 2008; Bengston et al. 2008; Alruily et al. 2009; Shaalan, K., & Raza, H. 2009; Riloff E. 2007; Alkaff 2012) have successfully guided analysts with crime cases (Hao et al. 2008).

In the domain of crime and crime analysis, information on crime is needed as quickly as possible and as accurately as possible. It is difficult to manually access data that is needed for investigations (Hao 2008).

Chao et al. (2002) created a neural network based entity extractor, which implements named entity extraction techniques to extract address, person, drugs and personal property from crime documents and police reports. The model had a precision value and recall value for person name of 74.1% and 73.4% respectively. The precision value and recall value for narcotic drugs was 85.4% and 77.9% respectively. The model performed with greater accuracy for narcotic drugs.

Hao et al. (2008) created an IE model customized for the crime domain that uses Natural Language Processing (NLP) to identify crime related information from police reports, witness narratives, and news documents. The model extracts people, vehicles, weapons, time, locations, and clothes, and was tested on two different types of documents, which were police narrative reports and witness narrative reports, that were all obtained from online forums, blogs, and news documents. The model uses both the lexical lookup and rule-based approaches. The model had achieved high precision and recall values of 96% and 83%, respectively, when tested on police narrative reports. However, the model achieved lower precision and recall values of 93% and 77% when tested on witness narrative reports.

Alruily et al. (2009) had created the Crime Type Recognition System (CTRS) that recognizes different crime types and uses two combined techniques. The first one is direct recognition using gazetteers of crime verbs and crime names. The second one is completely rule based, and relies on several rules and a crime indicator list to identify crime types. The work of Alruily et al. (2009) was developed based on the previous research of Shaalan and Raza (2009), who had used a rule based approach to create a named entity recognition system, and also based on the research of Poibeau (2003), who had developed a multilingual named entity recognition framework using the rule based approach. The model was tested against human based entity extraction, and the precision, recall and F-measure were recorded to be 60%, 97%, and 74% respectively. By increasing the value of the recall, the precision had decreased accordingly.

Alkaff (2012) had created an IE model that extracts the nationality from online crime news documents, and uses coreference identification to associate the nationality to either a suspect or victim or none (if nothing is matched). Regarding the evaluation of the direct and indirect extraction approaches, the precision, recall, and F-measure values were 55%, 96% and 70% respectively. Regarding the evaluation of the victim or suspect reference identification, the precision, recall, and F-measure values were 62%, 53%, and 57% respectively. Although the model was effective, the approach used is not dynamic because it references nationality to suspect or victim based on the nearest keyword from the position of the nationality. Hence, it is capable of being enhanced.

3. Proposed Model

The proposed model was developed using three main stages. The first stage involved the creation of the corpus. The second stage involved the generation of the gazetteers and internal lists. The third stage is about the implementation of the proposed model, which includes the model architecture, the components, and the techniques used for each component.

3.1 Stage 1 – Corpus Creation

During the first stage, data related to the crime domain was gathered to create the corpus used for this work. The data source used for the corpus was collected and gathered from Bernama, the Malaysian national news agency. The test corpus includes approximately 248 crime news documents, which have been stored on a local computer and used during the implementation of the model. Forty eight of the documents are from the same dataset used by Alkaff (2012).

3.2 Stage 2 – Generation of Gazetteers and Internal Lists

During the second stage, the gazetteers and internal lists were gathered. The nationality list (NL), nationality indicator list (NIL), and country list (CL) were collected from Wikipedia, and also from the GATE gazetteer lists. In addition, some manual additions were made to the NIL by searching the documents for any words that indicate a nationality, such as "from" and "national". These lists are used to check if there is a nationality match in the nationality extractor component. Figures 1, 2 and 3 show samples of the NL, NIL, and CL respectively.

from, national,
nationals, nationality,
nationalities, originally
origin.

Figure1. Nationality Indicator List

Koreans, Kosovo Albanians,
Kuwaitis, Lao, Latvians,
Lebanese, Liberians, Libyans,
Liechtensteiners, Lithuanians,
Luxembourgers, Macedonians,

Figure 2. Nationality List

French Guiana, French Polynesia, French Southern, and Antarctic Lands, Gabon, Gambia, Gambie, Gaza Strip, German Democratic Republic, Germany, Ghana, Gibraltar, Great Britain, Greece, Greenland, Grenada, Grenade, Groenland, Grèce, Guadeloupe, Guam, Guatemala, Guernesey, Guernsey, Guinea, Guyanem, Haiti, Holland, Honduras, Hong Kong, …

Figure 3. Country List

The victim keywords list (VKL) and suspect keywords list (SKL) consist of both verbs and nouns. Figure 4 and 5 show the VKL and the SKL, respectively.

victim, victims, dead,
died, hospitalized, wounded,
stabbed, suffer, suffered,

Figure 4. Victim Keywords List

suspect, suspects, caught,
nabbed, committed, detained,
detainees, arrested,

Figure 5. Suspect Keywords List

3.3 Stage 3 – Proposed Model Implementation

The proposed model uses a hybrid approach involving a lexical lookup approach and a rule based approach. It was implemented using the Java programming language. Figure 6 shows the proposed model, which contains the preprocessing component, the nationality extractor component, and the victim or suspect reference component. In addition, the figure also shows the gazetteers and lists used, and the inputs and outputs going to and from the model.

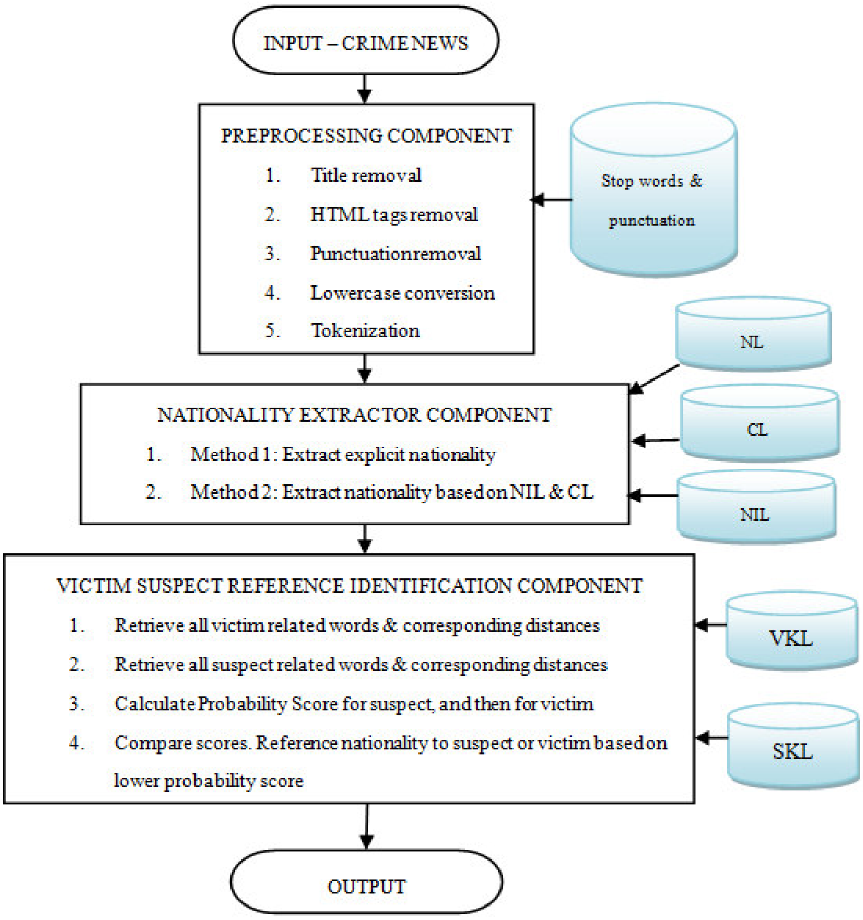

INPUT – CRIME NEWS
PREPROCESSING COMPONENT
1. Title removal
2. HTML tags removal
3. Punctuation removal
4. Lowercase conversion
5. Tokenization
Stop words & punctuation
NL
CL
NIL
NATIONALITY EXTRACTOR COMPONENT
1. Method 1: Extract explicit nationality
2. Method 2: Extract nationality based on NIL & CL
VICTIM SUSPECT REFERENCE IDENTIFICATION COMPONENT
1. Retrieve all victim related words & corresponding distances
2. Retrieve all suspect related words & corresponding distances
3. Calculate Probability Score for suspect, and then for victim
4. Compare scores. Reference nationality to suspect or victim based on lower probability score
VKL
SKL
OUTPUT

Figure 6. Proposed model architecture and components

Preprocessing Component

The preprocessing component processes the text before it is sent to the other components so that the text is ready for processing by the other components. This way, the process of extraction is more efficient, because the other components do not have to deal with any additional irrelevant tokens. The preprocessing component includes five main parts, which are title removal, HTML removal, punctuation removal, lowercase conversion, and tokenization.

Nationality Extractor Component

The next component is the nationality extractor component, which attempts to extract all nationalities from crime news documents. The output of this component is shown to the user, and is also used as input for the next component, which is the reference identification component. This component works based on two algorithms, the direct match algorithm and the indirect match algorithm. The model first uses the direct algorithm to check for matches using the nationality list (NL). An example of a direct match is “Indonesian”. The direct match algorithm is shown in Figure 7.

```
Require: $file \neq 0$
token <= file
nl <= NationalityList
while $file \neq 0$
    for nlCounter <= 0 to length[nl]-1   do
        if   token == nl[nlCounter]   then
            NationalityExtracted <= token
        end if
    end for
end while
```

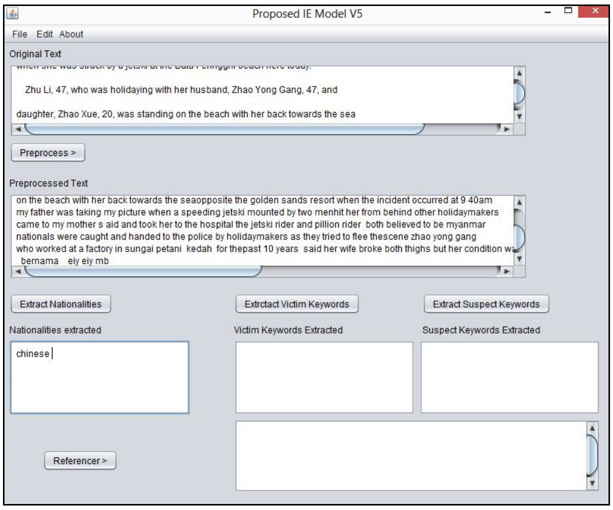
Figure 7. Direct match algorithm

If there are no matches found using the direct match algorithm, the component moves on to attempt to use the indirect match algorithm, along with the nationality indicator list (NIL) and the country list (CL) to find any matches. An example of an indirect match is “from Indonesia”. The word “from” is matched using the NIL, and the nationality “Indonesia” is matched using the CL. Figure 8 shows the indirect match algorithm.

```
Require: $file \neq 0$
token <= file
previousToken <= token[-1]
nextToken <= token[+1]
cl<=CountryList
nil<=NationalityIndicatorList
while $file \neq 0$
    fornilCounter <= 0 to length[NIL]-1   do
        if   token == nil[nilCounter]   then
            NationalityExtracted <= token
        end if
    end for
    for nilCounter <= 0 to length[nil]-1   do
        if   token == nl[nilCounter]   then
            for clCounter <=0  to length[cl]-1   do
                if  previousToken == cl[clCounter]  or
                    nextToken ==  cl[clCounter]   then
                    NationalityExtracted <= previousToken +token or
                    NationalityExtracted <= token +nextToken
                end if
            end for
        end if
    end for
end while
```

Figure 8. Indirect match algorithm

After this component has extracted all of the nationalities in the text, either by using the direct match algorithm or indirect match algorithm, it stores all of the nationalities in an array. This array of nationalities is output to the user, and also used as input to the reference identification component. Figure 9 shows the output of the nationality extractor component after it is processed on a sample text from the corpus.



when she was struck by a jetski at the Batu Ferringhi beach here today.

Zhu Li, 47, who was holidaying with her husband, Zhao Yong Gang, 47, and

daughter, Zhao Xue, 20, was standing on the beach with her back towards the sea

on the beach with her back towards the seaopposite the golden sands resort when the incident occurred at 9 40am
my father was taking my picture when a speeding jetski mounted by two menhit her from behind other holidaymakers
came to my mother s aid and took her to the hospital the jetski rider and pillion rider both believed to be myanmar
nationals were caught and handed to the police by holidaymakers as they tried to flee thescene zhao yong gang
who worked at a factory in sungai petani kedah for thepast 10 years said her wife broke both thighs but her condition wa
bernama eiy eiy mb

chinese

Figure 9. Output of nationality extractor component

As shown on the GUI of the model, when the user clicks on the “Extract Nationalities” button, the system extracts all of the nationalities from the text, and outputs them to the text area under “Nationalities extracted”. It is critical that this component outputs results correctly, as the next component, which is the reference identification component, depends on this output.

Reference Identification Component

Finally, the last component is the victim or suspect reference identification component. During model execution, for each nationality, this component attempts to associate the nationality extracted to being a suspect or victim. It does this using the victim keywords list (VKL) and suspect keywords list (SKL). First, it uses the VKL to find all of the words in the text that are victim related. Next, it calculates the distance of each victim related word in relation to the position of the nationality extracted. For example, if a victim related word is four words away from the nationality within the text, it has a distance of four (distance = 4.0).

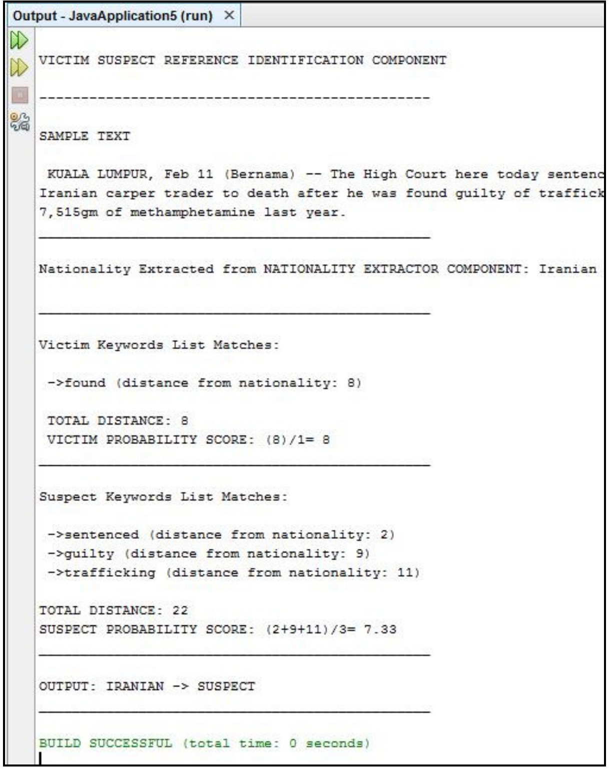
After that, the distances of all the victim related words are added to give the total distance, which is divided by their count in order to get the average distance. The average distance is the victim probability score. Next, the suspect probability score is calculated using the same technique used to calculate the victim probability score. These two probability scores are stored by this component to be used for comparison.

Both the victim probability score and the suspect probability score are compared. If the victim probability score is less than the suspect probability score, then the nationality extracted is referenced as that which belongs to the victim feature. This is because the victim probability score is less, which means that the average distance between the nationality position and the victim related keywords positions is less than the average distance between the nationality position and the suspect related keywords positions. Hence, the probability of the nationality being related to the victim is higher due to the shorter distance of the victim related keywords from the nationality position. Figure 10 shows the reference identification algorithm used by the model.

```
Require: file ≠ 0
while file ≠ 0
    find positions of all victim related keywords
    find positions of all suspect related keywords
    for n <= 0 to length[nationalityArray]-1 do
        for victimKeyword <= 0 to length[victimKeywordsArray]-1 do
            calculate distance between keyword and nationality
            totalDistance <= sum of distance of all keywords from nationality
        end for
        avgDistance <= totalDistance / countOfVictimKeywords
        victimProbabilityScore <= avgDistance
        for suspectKeyword <= 0 to length[suspectKeywordsArray]-1 do
            calculate distance between keyword and nationality
            totalDistance <= sum of distance of all keywords from nationality
        end for
        avgDistance <= totalDistance / countOfSuspectKeywords
        victimProbabilityScore <= avgDistance
        IF (victimProbabilityScore < suspectProbabilityScore)
            Nationality referenced to victim
        ELSE
            Nationality referenced to suspect
    end for
end while
```

Figure 10. Reference identification algorithm

Figure 11 shows the victim suspect reference identification component after processing a snippet of text from the corpus (Note that Figure 11 shows the console based output and not the GUI based output because the GUI based output is not able to be shown at full length, due to the small text area).



```
VICTIM SUSPECT REFERENCE IDENTIFICATION COMPONENT

---------------------------------------------

SAMPLE TEXT

 KUALA LUMPUR, Feb 11 (Bernama) -- The High Court here today sentenc
Iranian carper trader to death after he was found guilty of traffick
7,515gm of methamphetamine last year.
_____________________________________________

Nationality Extracted from NATIONALITY EXTRACTOR COMPONENT: Iranian

_____________________________________________

Victim Keywords List Matches:

 ->found (distance from nationality: 8)

 TOTAL DISTANCE: 8
 VICTIM PROBABILITY SCORE: (8)/1= 8
_____________________________________________

Suspect Keywords List Matches:

 ->sentenced (distance from nationality: 2)
 ->guilty (distance from nationality: 9)
 ->trafficking (distance from nationality: 11)

TOTAL DISTANCE: 22
SUSPECT PROBABILITY SCORE: (2+9+11)/3= 7.33
_____________________________________________

OUTPUT: IRANIAN -> SUSPECT
_____________________________________________

BUILD SUCCESSFUL (total time: 0 seconds)
```

Figure 11. Output of victim suspect reference identification component

In this case, the average distance between the victim related keywords and the nationality is 8. This is the victim probability score. The average distance between the suspect related keywords and the nationality is

7.33. This is the suspect probability score. The suspect probability score is less than the victim probability score, which means that the suspect related keywords are nearer to the position of the nationality, and therefore the nationality “Iranian” is referenced to “Suspect”.

The suspect reference identification component references all of the nationalities in the document to either the suspect or victim, based on the reference identification algorithm mentioned in Figure 10. The user clicks on the “Extract Victim Keywords” button in order for the model to extract the victim related keywords and display them in the text area under “Victim Keywords Extracted”. Next, the user clicks on the “Extract Suspect Keywords” button in order for the model to extract the suspect related keywords and display them in the text area under “Suspect Keywords Extracted”.

Finally, the user clicks on the “Referencer” button in order for the model to reference the nationality to the suspect or victim based on the reference identification algorithm.

4. Evaluation and Analysis of Results

This section presents the evaluation of the effectiveness and efficiency of the proposed IE model during the process of nationality extraction and reference identification to suspect or victim features. The model was used to process a total 248 crime news documents from the test data used in this work. 48 random documents were selected to be included in the experiments that were performed on the model.

Each document was input into the system, and went through the three system components (the preprocessing component, the nationality extractor component and the reference identification component) mentioned previously, in order to have the nationalities extracted, and to have each nationality referenced. In addition, all of the same 48 documents were processed manually, by the researcher. Both manual processing and system processing were compared in order to measure how well the model did in terms of the efficiency and accuracy. Two experiments were conducted. The first one was to evaluate the nationality extraction component of the model. The second one was to evaluate the reference identification component of the model.

The effectiveness, or accuracy, of the IE model was measured in terms of precision, recall and F-measure. These evaluation metrics are the standard metrics in use for measuring how well information extraction systems perform (Manning 2009). To do an evaluation on a model, it should contain a document collection, specific information to be extracted (such as person name) and a binary value to state whether the extracted piece of information is relevant or not relevant (Manning 2009; Cunningham H. 2006).

Experiment 1 – Evaluation of Nationality Extractor Component

The first experiment was conducted to evaluate the nationality extraction component of the model. For every document, the researcher had extracted all of the nationalities in the documents, after reading and comprehending the text.

Table 1 shows the nationalities that were extracted manually. Any document that did not have a nationality in it was denoted “none” under the “Nationalities extracted” column. The first column represents the document name, the second represents the nationalities extracted, and the third represents the number of nationalities in the document, respectively.

Table 1. List of nationalities extracted manually

Name	Nationalities Extracted	Nationalities
doc (1)	Indonesian	1
doc (2)	Indonesian, Indonesian	2
doc (3)	Indonesian, Indonesian	2
doc (4)	None	0
doc (5)	Malaysian, Indian, Malaysian, Malaysian, Indian, Malaysian	6
doc (6)	Chinese	1
doc (7)	Indonesian	1
doc (8)	Chinese	1
doc (9)	Chinese	1
doc (10)	None	0
doc (11)	Bangladeshi, Bangladeshi	2
doc (12)	Malaysians, Indonesian , Canadian, Canadian, Malaysian	5
doc (13)	Singaporean	1
doc (14)	Malaysian	1
doc (15)	Iranian	1
doc (16)	None	0
doc (17)	Bangladeshi	1
doc (18)	Malaysian	1
doc (19)	Malaysian, Thai	2
doc (20)	Malaysian, Indian, Malaysian, Malaysian, Indians, Malaysian	6
doc (21)	Malaysian, Iranian, Nigerian, Malaysians, Malaysians, Iranians	6
doc (22)	Thai, British, Malaysian	3
doc (23)	Australian	1
doc (24)	Iranian, Ugandan	2
doc (25)	Malaysian	1
doc (26)	None	0
doc (27)	Malaysian	1
doc (28)	Malaysian, Malaysian	2
doc (29)	Malaysian, Indian, Indian, Indian, Malaysian, Indians, Indian, Indian, Indian, Malaysian, Indian, Malaysian, Malaysian, Malaysian, Indian, Indian, Indian, Indian	18
doc (30)	Indonesian, Cambodian, Pakistani	3
doc (31)	Malaysians, Chinese, Australian, Malaysian	4
doc (32)	Chinese	1
doc (33)	Malaysian, Malaysian	2
doc (34)	Australian	1
doc (35)	Malaysian	1
doc (36)	Malaysian, Malaysian	2
doc (37)	Malaysian	1
doc (38)	Malaysian	1
doc (39)	Malaysian	1
doc (40)	Malaysian	1
doc (41)	Iranians, Iranian, Iranian, Iranians, Iranians	5
doc (42)	Chinese	1
doc (43)	Chinese	1
doc (44)	Malay, Indonesian, Indonesian	3
doc (45)	Malaysian	1
doc (46)	Malaysian, Malays, Indians	3
doc (47)	Malaysian	1
doc (48)	Malaysian, Malaysian	2

After extracting all of the nationalities from the test data, several points were inferred, which are shown in table 2.

Table 2. Information collected after nationality extraction

No. of Documents	No. of Correct Nationalities Extracted by Model	No. of Incorrect Nationalities Extracted by Model	Total no. of Nationalities
48	98	10	104

After calculating the precision, recall, and F-measure, the evaluation metrics for nationality extraction by the model are shown in table 3.

Table 3. Precision, recall and F-measure evaluation metrics

Evaluation Type	Precision	Recall	F-measure
Results	90%	94%	91%

Several points may be obtained from the results. The precision was 90%, which is relatively high. The system was able to extract 98 correct results and only 10 incorrect results. The recall was 94%, which was also relatively high. The system was able to extract 98 results correctly out of the total of 104. After calculating the precision and recall, their values were used as input for the F-measure formula, which resulted in 91%.

Experiment 2 – Evaluation of Reference Identification Component

The second experiment was conducted to evaluate the reference identification component of the model. In this experiment, the 104 nationalities that were manually extracted from the test data in the first experiment were used. For every document, the researcher had manually referenced every nationality, and recorded whether the nationality was associated with either the victim feature or the suspect feature, after reading and comprehending the text. After the reference identification process was done manually, it was performed by the proposed model.

Table 4 shows the nationalities, and the reference identification for each nationality, which were manually recorded. The third column is entitled References, and shows the reference value for each nationality. The reference values are S, for suspect, V, for victim, and N, for none (e.g., in "Chinese New Year").

The total of 104 extracted nationalities contained 57 nationalities that were not associated with victim or suspect features, and 47 that were. The model made 49 attempts in total to reference the 47 nationalities that were associated with features. The reason that the model made 49 attempts and only referenced 47 nationalities was because there were two failed attempts. There is an issue where the model was able to find victim related keywords only and not suspect related keywords (or the other way around), and therefore would not reference the nationality. When this issue occurs during execution, it is recorded as a failed attempt to reference the nationality. Only 15 nationalities were referenced by the system incorrectly, and the remaining 32 were referenced correctly, as shown in table 5.

Table 5. Data after reference identification component

No. of Reference Attempts	No. of Correct References	No. of Incorrect References
49	32	15

After calculating the precision, recall, and F-measure, the evaluation metrics for nationality extraction by the model are shown in table 6.

Table 6. Precision, recall and F-measure evaluation metrics

Evaluation	Precision	Recall	F-measure
Results	65%	68%	66%

Several points may be obtained from the results. The precision was 65%. The system was able to correctly reference 32 nationalities in 49 attempts. The recall was 68%, which only 2% higher than the precision. The system had 32 correct references out of the total 47 references. After calculating the precision and recall, their values were used as input for the F-measure formula, which resulted in 66%.

Analysis of Results

Several observations may be concluded from the results of the two experiments conducted on the model. Before presenting these observations, an overview and comparison of the results is shown.

Table 4. Manual reference identification of nationality

Name	Nationalities extracted	References
doc (1)	Indonesian	s
doc (2)	Indonesian, Indonesian	s, s
doc (3)	Indonesian, Indonesian	v, n
doc (4)	none	n
doc (5)	Malaysian, Indian, Malaysian, Malaysian, Indian, Malaysian	v, n, n, n, n, n
doc (6)	Chinese	v
doc (7)	Indonesian	v
doc (8)	Chinese	n
doc (9)	Chinese	v
doc (10)	None	n
doc (11)	Bangladeshi, Bangladeshi	v, v
doc (12)	Malaysians, Indonesian, Canadian, Canadian, Malaysian	s, s, n, n, n
doc (13)	Singaporean	v
doc (14)	Malaysian	n
doc (15)	Iranian	s
doc (16)	None	n
doc (17)	Bangladeshi	v
doc (18)	Malaysian	n
doc (19)	Malaysian, Thai	n, n
doc (20)	Malaysian, Indian, Malaysian, Malaysian, Indians, Malaysian	s, n, n, n, v, v
doc (21)	Malaysian, Iranian, Nigerian, Malaysians, Malaysians, Iranians	s, s, s, s, s, s
doc (22)	Thai, British, Malaysian	n, n, n
doc (23)	Australian	n
doc (24)	Iranian, Ugandan	s, s
doc (25)	Malaysian	n
doc (26)	None	n
doc (27)	Malaysian	n
doc (28)	Malaysian, Malaysian, Malaysian	n, n, n
doc (29)	Malaysian, Indian, Indian, Indian, Malaysian, Indians, Indian, Indian, Indian, Malaysian, Indian, Malaysian, Malaysian, Malaysian, Indian, Indian, Indian, Indian	s, n, s, s, s, v, n, n, s, s, s, v, n, n, n, n, n, n
doc (30)	Indonesian, Cambodian, Pakistani	s, s, s
doc (31)	Malaysians, Chinese, Australian, Malaysian	s, s, n, n
doc (32)	Chinese	n
doc (33)	Malaysian, Malaysian	n, n
doc (34)	Australian	n
doc (35)	Malaysian	n
doc (36)	Malaysian, Malaysian	n, n
doc (37)	Malaysian	n
doc (38)	Malaysian	n
doc (39)	Malaysian	n
doc (40)	Malaysian	n
doc (41)	Iranians, Iranian, Iranian, Iranians, Iranians	s, s, s, s, s
doc (42)	Chinese	s
doc (43)	Chinese	n
doc (44)	Malay, Indonesian, Indonesian	n, s, n
doc (45)	Malaysian	n
doc (46)	Malaysian, Malays, Indians	n, n, n
doc (47)	Malaysian	n
doc (48)	Malaysian, Malaysian	n, n

The results from the first experiment were promising. The precision was 90%, the recall was 94%, and the F-measure was 91%. Alkaff (2012)'s work was used as a main reference for this research. Taking a look at his nationality extractor component, its evaluation metrics were 55% for precision, 96% for recall, and 70% for F-measure. After the results from the extraction components of both models were compared, the proposed model achieved higher results. Overall, these results are relatively high and this component did its job successfully with high evaluation metrics.

The results from the second experiment showed that the reference component is promising, and performed well according to the evaluation metrics. The precision was 65%, the recall was 68%, and the F-measure was 66%. Alkaff (2012)'s reference identification component had a precision, recall, and F-measure value of 62%, 53%, and 57%, respectively. After the results from the reference identification components of both models were compared, the proposed model achieved relatively higher results.

5. Conclusion and Future Work

This work proposed a model that is used to extract the nationality feature from crime news documents. The process of extraction was relatively accurate, with 90% precision, 94% recall, and 91% F-measure evaluation metrics, according to the results shown in Experiment 1.

The proposed model is also used to perform coreference identification to associate the nationality to specific features, based on an enhanced probability algorithm. The coreference identification process was not very accurate, but outperformed other systems, with 65% precision, 68% recall, and 66% F-measure evaluation metrics, according to the results shown in Experiment 2.

This work has created a promising IE system that may be the foundation for future works related to this research area. The following suggestions may be used for future work:

1. The development of the model to successfully deal with crime news documents from companies other than Bernama.
2. The development of the model to process crime news documents in other languages, or, perhaps a multilingual model that accepts multiple languages.
3. The capability of the model to extract other useful crime related entities and features, as this is beneficial to crime analysts. Entities such as crime locations, crime dates, crime types, criminal properties, weapons used, narcotic drugs, car brand, among others, may be taken into consideration (Chau et al. 2002; Feldman et al. 2006).
4. The model uses the integration of the rule based approach and the lexical lookup based approach. It may be enhanced dramatically by integrating other approaches, such as semantic based, machine learning and neural network based approaches.

An IE model has been implemented with success, and was able to perform named entity extraction and coreference identification on crime news documents of an unstructured nature. This work has achieved promising results, and, in conclusion, is predicted to open a new path for future research related to information extraction in the crime domain.

References

Alruily, M., Aladdin, A. & Abdulsamad., A. (2010), "Using Self Organizing Map to cluster Arabic crime documents", *Proceedings of the International Multiconference on Computer Science and Information Technology*. IMCSIT, 357-363.

Alruily, M., Aladdin, A. & Zedan, H. (2009), "Crime Type Document Classification from Arabic Corpus", *Second International Conference on Developments in eSystems Engineering*. DESE, 153-159.

Alwee, R., Shamsuddin, S. & Sallehuddin, R. (2013), "Economic indicators selection for crime rates forecasting using cooperative feature selection", *AIP Conference Proceedings*, 1522(1): 1221-1231.

Appelt, D. (1999), "Introduction to information extraction", *AI Commun.* 12(3): 161-172.

Bao, S., Zhang, L., Chen, E., Long, M., Li, R. & Yu, Y. (2006), "LSM: Language Sense Model for Information Retrieval", *Advances in Web-Age Information Management*. 4016, Springer Berlin Heidelberg, 97-108.

Borthwick, A. (2009), "Co‐Reference in GATE", *Principal Scientist*.

Chau, M., Xu, J. & Chen, H. (2002), "Extracting meaningful entities from police narrative reports", *Anjuran Digital Government Society of North America*. Los Angeles, California, 2002: 1-5.

Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W. & Schroeder, J. (2003), "COPLINK: managing law enforcement data and knowledge", *Communications of the ACM*, 46(1): 28-34.

Hao, C., Iriberri, A. & Leroy, G. (2008), "Crime Information Extraction from Police and Witness Narrative Reports", *Conference on Technologies for Homeland Security*. 2008. IEEE, 193-198.

Crestani. F. (2008), "From Linking Text to Linking Crimes: Information Retrieval, But Not As You Know It", *Information Access through Search Engines and Digital Libraries* 22. Springer Berlin Heidelberg, pp. 13-42.

Cunningham, H. (2013), "Developing Language Processing Components With Gate Version 7", *The University of Sheffield*.

Cunningham, H. (2006), "Automatic Information Extraction", *Oxford: Elsevier volume 5*.

Han, X. & Zhao, J. (2009), "CASIANED: People Attribute Extraction based on Information Extraction", Anjuran Madrid, Spain.

Manning, C., Raghavan, P. & Schütze, H. (2008), "Introduction to information retrieval", Vol. 1. Cambridge: Cambridge University Press.

Pinheiro, V., Furtado, V. Pequeno T. & Nogueira, D. (2010), "Natural Language Processing based on Semantic inferentialism for extracting crime information from text", *IEEE International Conference on Intelligence and Security Informatics*. ISI. 2010. 19-24.

Ponte, J. &. Croft, B. (1998), "A language modeling approach to information retrieval". *Anjuran ACM*. Melbourne, Australia.

Riloff, E. & Lorenzen, J. (1999), "Extraction-Based Text Categorization: Generating Domain-Specific Role Relationships Automatically", *Natural Language Information Retrieval 7*. 167-196. Springer Netherlands.

Sidhu, S. (2006), "Crime Levels And Trends In The Next Decade". *Journal of the Kuala Lumpur Royal Malaysia Police College*.

Sundheim, B. (1991), "Overview of the third message understanding evaluation and conference", *Anjuran Association for Computational Linguistics*. San Diego, California.

Tin, C., Cua, J., Tan, M., Yao, K. & Roxas, R. (2009), "Information extraction from legal documents", *Natural Language Processing, 2009*. SNLP. Eighth International Symposium on. 157-162.

Williams D. & Poulovassilis, A. (2008), "Combining Data Integration and IE Techniques to Support Partially Structured Data". *Anjuran Springer-Verlag*. London, UK.

Wimalasuriya, D. (2010), "Components for Information Extraction: Ontology Based Information Extractors and Generic Platforms". *ACM*.

Wu, F. & Weld, D. (2010), "Open information extraction using Wikipedia". *Anjuran Association for Computational Linguistics*. Uppsala, Sweden.

Xiao, L., Wissmann, D., Brown, M. & Jablonski, S. (2004), "Information Extraction from the Web: System and Techniques". *Applied Intelligence 21*(2): 195-224.

Zhao, S. (2005), "Information extraction from multiple syntactic sources" *Tesis New York University*.