Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

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

EXTRACTION OF NATIONALITY FROM CRIME NEWS

[1]ABDULRAHMAN ALKAFF, [2]MASNIZAH MOHD
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia
E-mail: [1]walkaffw@gmail.com, [2]mas@ftsm.ukm.my

ABSTRACT

Most of the crimes committed today are reported on the Internet through news, blogs and social networking sites. These sources have provided a huge amount of crime data, presenting a need for a means to extract useful information. In this research, the evaluation of Direct and Indirect extraction of nationality from crime news, along with the additional references to identify the nationalities of suspects, victims and witnesses is presented. Named entity recognition using gazetteers and rule-based extraction are used, in addition to co-reference resolution to link references. The proposed approach was evaluated and compared to manual extraction system. The results indicate that the proposed approach is able to extract most information related to nationality from crimes news and identify the additional information references. Its performance proved good, with 55% precision, 96% recall, and 70% f-measure evaluation metrics.

Keywords: *Information Extraction, Crime, Nationality, Victim, Suspect*

1. INTRODUCTION

The amount of information available throughout the world is increasing rapidly. Content available to public users, policeman and expert analysts through various types of media includes both private data, exemplified by Wikileaks, and public data, such as crime news sources. Therefore, accessing information is not the problem; the challenge is how to retrieve and extract useful and relevant data concerning the specific user needs.

One technique employed to overcome this challenge is an information retrieval field, which means finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored in computers) [1].

A type of information retrieval called Information Extraction (IE) aims to automatically extract structured information such as entities, attributes of those entities, relations between entities and events from unstructured and/or semi-structured machine-readable documents [2].

In the crime domain, computing techniques are used to analyze crime data for investigation, monitoring, collecting and connecting the data. Information retrieval and extraction has particularly proven efficiency in solving and preventing crimes, both volume crimes (burglary, vehicle crime, etc.) and major crimes (fraud, drug trafficking, murder, etc.) [3-8].

People's names, organization's names, locations of crimes, personal attributes (e.g. age of a person), weapons used and nationalities are the most important entities in the crime domain [3].

Still, crime in Malaysia (that consists of many different nationalities, with many foreign tourists, immigrant workers, and international students) has increased tremendously over the past years [9-11]. Crime news reported in many sources provides information pertinent to crime data analysis. Therefore, this research aims at extracting useful entities from crime news, in particular at extracting and recognizing nationality of persons involved in criminal activities, and identifying whether or not nationality refers to suspects, victims, and witnesses.

2. RELATED WORKS

2.1 Background

Information retrieval and information extraction are well known. Surveys show that 85% of Internet users use such techniques and employ search engines as to locate information [12-13].

The huge amount of data available today on world news through the media (newspapers, radio, etc.) creates a need for tools that manipulate and analyse data. However, current techniques do not perform at levels as high as human performance [14-16].

Increasing amounts of information and the need to search for specific information as "the search for needle". Information retrieval systems provide a list of relevant documents, often very large, requiring each document to be examined for real relevance.

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

Information extraction technologies can help determine the authentic relevance of retrieved data by recognizing entities and identifying the relationships between those entities.

Due to the complexity of information and the difficulty of accessing relevant information, there is a need to retrieve and manipulate user's requirements quickly and efficiently. In advanced cases, there is a need to analyse the retrieved data as input for other techniques that are needed for more advanced analysing; yielding data in structured rather than unstructured formats [15], [17].

There are several information extraction systems related to specific domains. Information and computer technology, involving information retrieval and extraction have proven efficacy in solving and preventing crimes [8]. Some of them are of particular importance because of abilities in collecting and connecting data, investigations, future crime prevention and decision-making processes [18-20].

2.2 Crime Related Information Extraction Systems

Internet web sites, blogs, and forums are considered an efficient way to report crimes due to accessibility. In many cases, people like to record information about crimes in an informal way, without the formal complex forms and reports used by police [8], [21].

Currently, one of the most important applications of information extraction technology is intelligence gathering in the crime analysis domain. After the September 11 attacks, many information extraction experts helped police and intelligence services to find and link information available, in both private sources such as police reports and public sources such as news, chat rooms and web texts [22].

Extracting information in this domain is still limited because of the difficulty of obtaining the data. The Message Understanding Conferences MUC-3 and MUC-4 began applying information extraction on news articles regarding Latin American terrorism. During the Message Understanding Conferences (MUC), many extraction tasks were achieved, such as named entity recognition, organizations, locations, date, time, money and percentages. Other tasks are event extraction, their participants and settings, and scenario extraction, linking of individual events in a story line. Co-reference resolution determines whether two phrases refer to the same entity, person, time, place, and event in the world [15], [23].

Recently DARPA (Defence Advanced Research Projects Agency) started the EELD (Evidence Extraction and Link Discovery) program that aims to detect crime patterns, gangs, and suspect activities [7-8], [15]. However, MUC-3, MUC-4 and ACE ignored nationality of crime members. For their purposes, this entity was unimportant to their requirements. Hence, this paper aims to extract nationalities of involved persons in crimes.

[8] developed an information extraction system able to extract related-crime entities from sources by developing a lexicon combined with a rule-based system. In total, they built about 126 lexicon gazetteer lists. Despite the fact that this system used 126 lexicon gazetteers, a nationality gazetteer was not included. Hence, for a multinational community or international crime news sources, nationality extraction and the Victim Suspect-Identification approach proposed by this research might be useful in such systems to provide efficient extraction of crime related information.

In crime analysis, it is difficult to access data useful for intelligence investigations. [21] worked on a neural network-based entity extractor that used named-entity techniques to identify and extract meaningful entities from police narrative reports. Because the data becomes structured, it can be helpful in classification, such as criminal relationship identification and crime pattern recognition. Features that allow for predicting next crime locations for serial crimes, prioritizing suspects, and detecting general crime trends, can help in decision-making [15], [17].

[25] developed a system that extracts specific information within the crime domain from Arabic news articles corpus based on two recognition approaches: Direct recognition using gazetteers and rule-based recognition. In the first approach, they used a manually defined list of crime verbs (e.g. kill, steal, rape) and crime names (e.g. killer, robber, murder, drug) to recognize crime type entity. The other approach depended on a predefined indicator list used to indicate the type of crime indirectly by using indicators (e.g. perform, involve). However, a similar idea was implemented in this research, in particular the crime indicator list to identify the crime. Hence, this research is going to use this idea to indicate the nationality using Nationality Indicator Keywords List (NIKL) and Victim-Suspect Indicator Keywords List (VSIKL) to indicate or identify the suspects or victims.

Recently [26] developed a system capable of recognizing and extracting crime entities such as

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

types of crime, locations and nationalities from Arabic crime news sources taken from five different Arabic news agency websites. For nationality extraction, the system has some predefined Arabic keyword "nationality" to recognize the nationality stated before or after that keyword depending on Arabic rule of "the" occurrence.

However, this system used one nationality indicator to construct the nationality, which might be useful for Arabic crime news. Moreover, they used data customized from different sources. The same indicator idea is used in this research, but with additional roles and extra nationality descriptions.

[21], [27] used noun phrasing, lexical lookup, and the neural network methods to extract the entities of person, address, narcotic drug, and personal property in the proposal system. [22] presented a feature-based approach to extract relations from Thai news documents by applying a pattern-based named entity to extract people's names, locations and action entities.

[32] presented a new approach for information extraction pattern learning, aimed at benefiting from role-identifying nouns, role-identifying verb and role-identifying expressions. This approach used nouns, verbs, or phrases for role-identifying, meaning helping to reveal the role that the entity or object plays on an event. That approach was beneficial for information extraction, consequently, a similar idea is required to identify the nationality extracted by this research to differentiate victims, suspects and witnesses. However instead of identifying roles, it will be used to identify the nationality using nouns and verbs, so-called indicator keywords, depending on their lexical meaning.

[30] developed an anonymous Online Crime Reporting System to extract relevant crime information from witness narratives to collect more accurate and complete information about the crime that might help the police to prevent and solve the crimes. However, this system is limited to extracting crime related information directly from witnesses; information regarding nationality is not included. Hence, the nationality extraction module developed by this research might be useful to increase the performance of such systems.

[33] developed an information extraction system, called WikiCrimes IE, to extract crime information from texts and web sources to be used as input for collaborative web-based system of registering crimes called WikiCrimes (www.wikicrimes.org). However, extracting all crime related information is useful to build such web-based systems, and the nationality extraction module proposed by this research is one of those pieces of information.

2.3 Evaluation (Measures of Performance)

Information extraction has many standard evaluation metrics such as precision, recall and F-measure, which are the same metrics used in information retrieval.

[31] defined precision as a measure illustrates the accuracy of the system, which calculates the number of correct answers extracted by the system to the all extracted answers. The recall measure is the number of correct answers given by system to the total number of possible correct answers in the text. Finally, F-measure considers the balance between recall and precision using a parameter β, which always takes the value one when recall and precision have the same priority and β is greater than one when the precision is preferred and less than one when recall is preferred.

The first quantitative performance evaluation of Information Extraction technologies is sponsored by DARPA in 1991 for the Third Message Understanding Conference (MUC3) participants. The evaluation used 100 documents for every MUC3 participant system based on a blind test. All documents are hand-coded by humans to capture all relevant information. The results obtained are compared to the template generated by the system using the score program. In addition, the Message Understanding Conference MUC4 is evaluated using a similar performance evaluation to MUC3, but instead uses two test sets of 100 novel texts instead of one [24].

[22] evaluated their entity extractor by selecting a test bed randomly, containing 36 reports from the Phoenix Police Department database, tested the system by humans manually extracting the entities, and then using precision and recall as performance metrics.

[7] evaluated their crime information system by collecting a corpus from different sources of blogs, forums and websites provided by police departments. Then they randomly selected 20 police narratives and 20 witness narratives. Next they compared them with the system output by using the same precision and recall metrics.

[26] achieved good results on their system. The nationality dictionary builder recognized 118 correct nationalities out of 123 after removing some noisy data from the corpus. Rates of 81.94, 95.93 and 88.38 were earned for precision, recall and F-

Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

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

measure, respectively. However, good results achieved by this system were associated with customized data test sets collected from sources, whereas in this research, test set crime data for one month was selected as blind test set.

However, the same evaluation metrics applied for all systems shown above were employed to evaluate the nationality and their references extraction approach of this research.

3. DATA DESCRIPTION

The data used in this research was collected from the Malaysian National News Agency (BERNAMA) and classified manually. Each file consists of an article by a journalist about one crime or more and is represented in an html format file. A random training set was taken from this corpus to be used in the training stage for internal lists manual keyword extraction. Another complete set from a one-month period (February 2011), including all news provided by BERNAMA, was used in experiments to validate and evaluate this system. This experiment set consists of 206 crime news articles.

4. PRE-PROCESSING

4.1 Internal And External Lists

The main input data collected for this system from external sources such as the Nationality gazetteer, the Country gazetteer, Stopwords and punctuations were collected from Natural Language ToolKit NLTK[1].

Initially a small training corpus was randomly selected from the crime news corpus for use in the manual extraction training stage. The main purpose of this stage is to manually extract and build three main gazetteer lists: Extra Nationality Keywords List (ENKL), Nationality Indicator Keywords List (NIKL), and Victim-Suspect Indicator Keywords List (VSIKL).

After manually pre-processing and tokenization, removing repetition and reading all the selected articles, internal gazetteer lists were generated for use in the automatic extraction stage. Extra Nationality Keywords List (ENKL): includes all other nationalities not stated in nationality list, which is a general nationality description (e.g. local, foreigner, European, Asian, citizen).

Nationality Indicator Keywords List (NIKL): This list describes keywords that are not nationalities, but if preceded or followed by a country name, indicate a description of nationality (e.g. national, nationality).

Victim-Suspect Indicator Keywords List (VSIKL): keywords that can indicate or identify the suspects or victims, depending on how the journalist chooses to express the person's relationship to the crime (e.g. suspect, victim, arrested, witness, committed, implicated).

4.2 Pre-Processing and Filtering

To improve the process of extraction and results and reduce the processing time, the data of the crime news corpus needs to prepared and filtered by some initial processing steps such as Title Removing, HTML Tags Removing, Stopwords, Punctuations Removing, and Tokenization process.

5. NATIONALITY INFORMATION EXTRACTION SYSTEM

A typical information extraction system was specialized to generate a system that extracts the nationality entity from crime news. Figure 1 depicts the overall processes and components of that system. This specialized system consists of five main components: the manual extraction training stage for building internal lists, the pre-processing stage, the extraction and reference identification stage, collecting of external gazetteers, and generating the output structured information.

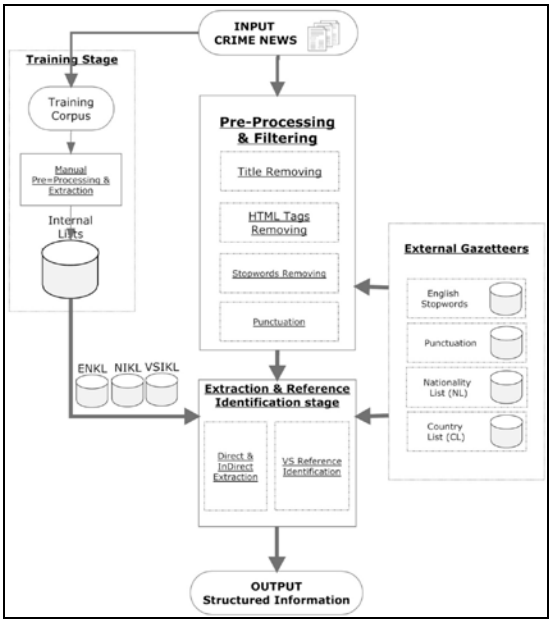

Figure 1: An overall Nationality Information Extraction System

[1] Free download from: www.nltk.org

Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

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

The main stage of this research comes after automatic pre-processing of all input crime news. This stage uses two main inputs, internal and external. An internal gazetteer list was generated from manual extraction during the training stage. These are an extra nationality keywords list (ENKL), a nationality indicator keywords list (NIKL), and a victim-suspect indicator keywords list (VSIKL).The second inputs are external gazetteers, which are the nationality list (NL) and the country list (CL). Below is an explanation of the architecture of this stage and its main components and modules.

The extraction and reference identification stage shows two main parts of the system: Extraction, including Direct and Indirect extraction, and identification of references to victims or suspects.

This stage is processed after tokenization pre-processing. Then, the system decides if it will execute the extraction modules or reference module depending on a Boolean value. This decision prevents the system from identifying a reference without a nationality. In this way, the system will allow the reference identification module only if there are nationalities extracted beforehand.

In the extraction module, the system starts recognizing nationalities using the Direct extraction algorithm, which is commonly used to express nationality. If nationality is not explicitly defined, the system will recognize it using the second approach, Indirect extraction. Otherwise, it will continue looking for references to nationality using tokens.

The next section describes the three modules and their algorithms and components.

5.1 Direct Extraction

The Direct extraction part of the system (*Algorithm 1*) is responsible for extracting the explicit nationality stated on the article, which occurs commonly. This part needs to access the external nationality gazetteer list (NL) then makes a comparison to all tokens that match the list. Extracted entities are stored in xml format.

Algorithm 1 Direct Extraction

```
    Require: file ≠ 0
1   token <= file
2   nl <= NationalityList
3   while file ≠ 0
4       for nlCounter <= 0 to length[nl]-1 do
5           if token == nl[nlCounter] then
6   NationalityExtracted <= token
7               end if
8       end for
9   end while
```

5.2 Indirect Extraction

If nationality cannot be derived by Direct extraction, Indirect Extraction (*Algorithm 2*) is responsible for recognizing Indirect ones. It needs to access three gazetteer lists: The Extra Nationality Keywords List (ENKL), the Nationality Indicator Keywords List (NIKL) and the Country gazetteer List (CL). Here, the system looks for matches between read tokens. All indirect nationalities listed in ENKL are first checked for. If not found, the system checks to see if there are any matches to the nationality indicator keywords list NIKL preceded or followed (depends on indicator keywords) by a country name. The system extracts the indicator keyword with the associated country name.

Algorithm 2 Rule-Based Extraction

```
    Require: file ≠ 0
1   token <= file
2   previousToken <= token[-1]
3   nextToken <= token[+1]
4   cl<=CountryList
5   enkl <= ExtraNationalityKeywordsList
6   nikl<=NationalityIndicatorKeywordsList
7   while file ≠ 0
8           forenklCounter <= 0 to length[enkl]-1 do
9             if token == enkl[enklCounter] then
10                NationalityExtracted <= token
11            end if
12  end for
13          for niklCounter <= 0 to length[nikl]-1 do
14              if token == nl[niklCounter] then
15           for clCounter <=0 to length[cl]-1 do
16           if previousToken == cl[clCounter] or
17             nextToken == cl[clCounter] then
18      NationalityExtracted <= previousToken +tokenor
19      NationalityExtracted <= token +nextToken
20                                  end if
21                              end for
22                      end if
23          end for
24  end while
```

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

5.3 Victim-Suspect Reference Identification

This approach matches the nationalities referred to as belonging to a suspect, a victim or witnesses. Most of journalists start by mentioning the name of the victim or the suspect, followed by information often including keywords without repeating the specific information (Algorithm 3).

In addition, Algorithm 3 does not execute only if there are nationalities extracted before by the previous parts. It needs to access an internal list from the Victim-Suspect Indicator Keywords List (VSIKL).

Algorithm 3 Victim-Suspect Reference Identification

```
    Require: file ≠ 0
1   token <= file
2   ne <= NationalityExtracted
3   vsikl<=VictimSuspectIndicatorKeywordsList
4   while file ≠ 0 and ne not changed
5   forvsiklCounter <= 0 to length[vsikl]-1 do
6   if token == vsikl[vsiklCounter] then
7   ne[identify] <= token
8           end if
9           end for
10  end while
```

6. EXPERIMENTAL RESULTS AND EVALUATION

This research was done both manually and through the system; results were compared to evaluate the effectiveness of the system.

The data set used for the evaluation stage covered one month (February 2011) of crime news articles provided by Bernama, consisting of 206 articles related to crime. About 70 of these files mentioned information regarding nationality.

6.1 Experiments

Four experiments were carried out to evaluate the system. The first experiment involved applying the Direct extraction approach for both manual and automatic systems. The second experiment was the Indirect extraction approach. Both Direct and Indirect approaches were applied for the third experiment. Finally, Victim-Suspect reference identification was covered in the last experiment.

Table 1 shows the results of all experiments. **No. of correct** is refers to the total numbers of correct answers extracted by the system. **No. of incorrect** is refers to the number of incorrect answers extracted by the system. Finally, **Goal** is refers to the number of all goal answers that should have extracted.

Table 1: Results of all experiments

	No. of correct	No. of incorrect	Goal
First Experiment	55	64	56
Second Experiment	38	12	41
Third Experiment	93	76	97
Fourth Experiment	51	31	96

6.2 Evaluation

The most common performance metrics used for information extraction are precision, recall and F-measure. Precision measure is used to measure the accuracy of the information extraction system as:

$$Precision = \frac{number\ of\ entities\ extracted\ correctly\ by\ the\ system}{all\ entities\ extracted\ by\ system}$$

Recall measure is defined as the percentage of relevant information extracted by the system:

$$Recall = \frac{number\ of\ entities\ extracted\ correctly\ by\ the\ system}{all\ entities\ extracted\ by\ manual\ system}$$

The last measure, called F-measure, combines precision and recall into a single measurement to balance them.

$$F - measure = \frac{2PR}{P + R}$$

Here, P stands for precision and R stands for recall. Table 2 shows the evaluation results of all experiments.

Table 2: Evaluation results of all experiments

	Precision	Recall	F-measure
First Experiment	46%	98%	63%
Second Experiment	76%	93%	84%
Third Experiment	55%	96%	70%
Fourth Experiment	62%	53%	57%

The experiments outlined show precision and recall relations, as presented in Figure 2 The four lines here represent the average: line for Direct extraction, Indirect extraction, both Direct and Indirect extraction and victim-suspect reference identification.

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

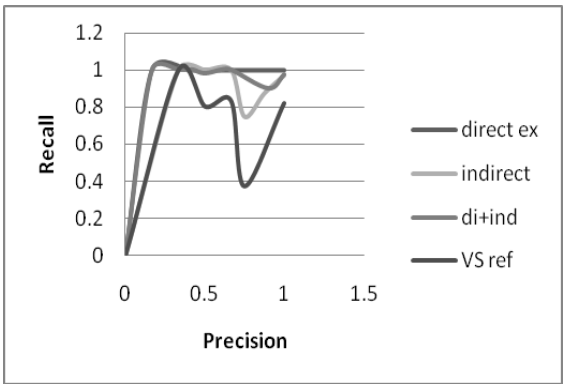

Figure.2: Precision And Recall For Nationality Extraction System Stages

This figure clearly shows that the system is able to recognize and extract nationality in an efficient way. The performance of the system in the identification of victim-suspect references is less consistent, but does not altogether fail.

6.3 Result Analysis

Analysing the above results, the research notes the following points:

- Recall of Direct Extraction is very high, meaning the system extracted most of direct nationalities that should be extracted. Some nationalities are not listed in the NLTK nationality gazetteer list, such as Myanmar.
- Precision of Direct Extraction was not high. Many incorrect nationalities were extracted by the system, which might be the result of including text such as Malaysian Anti Corporation, Chinese New Year and Australian Agency News, which were commonly used. A suggestion for future work is to extract instances such as these organization names.
- Using punctuation instead of nationality indicators led the system to identify only the nationality that followed the indicator, such as "from Indonesia, Cambodia or Pakistan". This led the system to extract "from Indonesian" only and ignore other country references. More rules are needed to solve this problem and are suggested in future work.
- The system used the country list imported from NLTK toolkit, but it did not include all countries, including Moldova and Cambodia. An updated and comprehensive list needs to be developed.
- The reporter may identify the state of victims or suspects in implicit way, not defined by keywords. Although understood by human as common style writing, the system was not effective at picking up these markers. It might be possible to improve this process by semantic extraction, suggested for future work.
- The order of sentences and the manner of expression may cause the system to refer to incorrect references or miss the references. This could be solved by including both anaphoric and cataphoric references and adding more rules, as suggested for future work.

7. CONCLUSION

In this paper, a system to extract nationalities in the crime domain from news provided by Bernama news agency is presented. This system consists of three modules: Direct extraction, Indirect extraction and a Victim-Suspect reference identification module.

The developed system was evaluated and tested by using a sample of news articles. The obtained results were compared to actual results, found through human manual extraction. In addition, performance metrics were calculated and the results were discussed and analysed in detail.

In the future, this system will be developed to be able to extract more crime related data from crime news, and serve as a useful tool for creating structured data for advanced analysis, investigations and retrieval queries.

REFRENCES:

[1] Manning, C.D., Raghavan, P. and Schutze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[2] Sunita, S., "Information Extraction", *Foundations and Trends in Databases*, Now Publishers Inc., Vol. 1, No. 3, 2008, pp: 261-377.

[3] Bache, R., Crestani, F., Canter, D. and Youngs, D., "Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes", *Third International Symposium on Information Assurance and Security*, 2007. IAS 2007, pp. 399-404.

[4] Goh, D., Cao, T., Sølvberg, I., Rasmussen, E., Bache, R., Crestani, F., Canter, D. and Youngs, D., "Mining Police Digital Archives to Link Criminal Styles with Offender Characteristics", *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, Springer Berlin, Heidelberg, 2007, pp. 493-494.

[5] Bache, R. and Crestani, F., "Towards an Automated Approach to Offender Profiling.

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

Computational Sciences and Its Applications", *International Conference on Computational Sciences and Its Applications, 2008. ICCSA* '08, pp. 537-545

[6] Bache, R. and Crestani, F., "An approach to indexing and clustering news stories using continuous language models", *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*, Springer-Verlag, 2010, pp. 109-116.

[7] Chih Hao, K., Alicia, I. and Gondy, L., "Natural language processing and e-Government: crime information extraction from heterogeneous data sources", *Proceedings of the 2008 international conference on Digital government research*. Montreal, Canada: Digital Government Society of North America, 2008, pp. 162-170.

[8] Chih Hao, K., Iriberri, A. and Leroy, G. (2008), "Crime Information Extraction from Police and Witness Narrative Reports", *Conference on Technologies for Homeland Security*, 2008 IEEE, pp. 193-198.

[9] Sidhu, Amar Singh., "The rise of crime in Malaysia: an academic and statistical analysis", *Journal of the Kuala Lumpur Royal Malaysia Police College*. No. 4, 2005.

[10] Sidhu, Amar Singh., "Crime levels and trends in the next decade", *Journal of the Kuala Lumpur Royal Malaysia Police College*. No. 5, 2006.

[11] Tang, C.F., "The Linkages among Inflation, Unemployment and Crime Rates in Malaysia", *Int. Journal of Economics and Management*, Vol. 3, No.1, 2009, pp. 50-61.

[12] Lawrence, S. and Giles, C.L., "Accessibility of information on the web", *Nature*, Vol.400, 1999, pp. 107–109.

[13] Keane, M.T., O'Brien, M. and Smyth, B., "Are people biased in their use of search engines?", *Communications of the ACM.*, Vol. 51, No. 2, 2008, pp. 49-52.

[14] Pazienza, M. and Grishman, R., "Information extraction: Techniques and challenges", *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, Springer Berlin, Heidelberg, 1997, pp. 10-27.

[15] Moens, M.F., *Information Extraction: Algorithms and Prospects in a Retrieval Context*, Springer-Verlag, New York, 2006.

[16] Han, X. and Zhao, J., "CASIANED: People Attribute Extraction based on Information Extraction", *City*, 2009, pp. 20-24.

[17] Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H. and Chen, H., "Using Coplink to analyze criminal justice data", *IEEE Computer: Computer Society*, Vol. 35, No. 3, 2002, pp. 30-37.

[18] Mehrotra, S., Zeng, D., Chen, H., Thuraisingham, B., Wang, Fei-Yue, Kumar, N., De Beer, J., Vanthienen, J. and Moens, M.F., "Intelligent Information Retrieval Tools for Police", *Intelligence and Security Informatics*, Springer Berlin, Heidelberg, 2006, pp. 664-665.

[19] Xiao, L., Wissmann, D., Brown, M. and Jablonski, S., "Information Extraction from the Web: System and Techniques", *Applied Intelligence*, Springer, Vol. 21, No. 2, 2004, pp. 195-224.

[20] Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C., Moens, M. and Hiemstra, D., "Information Extraction and Linking in a Retrieval Context", *Advances in Information Retrieval*, Springer Berlin, Heidelberg, 2009, pp. 810-813.

[21] Chau, M., Xu, J.J. and Chen, H., "Extracting meaningful entities from police narrative reports", *Proceedings of the 2002 annual national conference on Digital government research*. Los Angeles, California, Digital Government Society of North America, 2002, dg.o '02, pp. 1-5.

[22] Chen, H., Chung, W., Yi Qin, Chau, M., Xu, J.J., Wang, G., Zheng, R. and Atabakhsh, H., "Crime data mining: an overview and case studies", *Proceedings of the 2003 annual national conference on Digital government research*. Boston, MA: Digital Government Society of North America, 2003, pp. 1-5.

[23] Chinchor, N., Lewis, D.D. and Hirschman, L., "Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3)", *Computational Linguistics*, MIT Press., Vol.19, No. 3, 1993, pp. 409-449.

[24] Lehnert, W., Cardie, C., Fisher, D., McCarthy, J. and Riloff, E., "Evaluating an Information Extraction System", *Journal of Integrated Computer-Aided Engineering.*, Vol. 1, No. 6, 1994.

[25] Alruily, M., Ayesh, A. and Zedan, H., "Crime Type Document Classification from Arabic Corpus", *Second International Conference on*

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

Developments in eSystems Engineering (DESE), 2009, pp. 153-159

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

[27] Appelt, D. E. and Israel, D. J., *Introduction to Information Extraction Technology*. Stockholm, Schweden., 1999, http://www.ai.sri.com/\~{}appelt/ie-tutorial/IJCAI99.pdf [April 2011].

[28] Chen, H., Yang, C., Chau, M., Li, Shu-Hsing, Tongtep, N. and Theeramunkong, T., "A Feature-Based Approach for Relation Extraction from Thai News Documents", *Intelligence and Security Informatics*, Springer Berlin, Heidelberg, 2009, pp. 149-154.

[29] Feldman, R. and Sanger, J., "Information extraction", *The Text Mining Handbook: Advanced Approches in Analyzing Unstructured Data*, Cambridge university press, 2007, pp. 94-130.

[30] Iriberri, A. and Leroy, G., "Natural Language Processing and e-Government: Extracting Reusable Crime Report Information", *IEEE International Conference on Information Reuse and Integration, IRI 2007*, 2007, pp. 221-226.

[31] Jurafsky, D. and Martin, J.H., *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, New Jersey, 2000.

[32] Phillips, W. and Riloff, E., "Exploiting Role-Identifying Nouns and Expressions for Information Extraction", *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP-07)*, 2007, pp. 165-172.

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