Journal of Theoretical and Applied Information Technology
15th October 2012. Vol. 44 No.1

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

ARABIC NAMED ENTITY RECOGNITION IN CRIME DOCUMENTS

M. ASHAREF[1], N. OMAR[2], M. ALBARED[3]
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Malaysia
E-mail: [1]mmaasharef@yahoo.com, [2]no@ftsm.ukm.my,[3] mohammed_albared@yahoo.com

ABSTRACT

Named entity recognition (NER) systems aim to automatically identify and classify the proper nouns in text. NER systems play a significant role in many areas of Natural Language Processing (NLP) such as question answering systems, text summarization and information retrieval. Unlike previous Arabic NER systems which have been built to extract named entities from general Arabic text, our task involves extracting named entities from crime documents. Extracting named entities from crime text provides basic information for crime analysis. This paper presents a rule-based approach to Arabic NER system relevant to the crime domain. Based on morphological information, predefined crime and general indicator lists and an Arabic named entity annotation corpus from crime domain, several syntactical rules and patterns of Arabic NER are induced and then formalized. Then, these rules and patterns are applied to identify and classify named entities in Arabic crime text. The result shows that the accuracy of our system is 90%, and this result indicates that the method is effective and the performance of the achieved system is satisfactory.

Keywords: *Natural Language Processing, Named Entity Recognition, Arabic Crime Documents.*

1. INTRODUCTION

With the rapid increase of the crime rate in the Arabic world and the volume of the crime information that is available on the web, this make the process of analyzing and finding relevant and in time information such as named entities from these crime documents is very important. For example, the Almotawaset online newspaper has published that in Algeria only, during March 2011, more than 4 thousand crime cases has been recorded. Moreover, recognition and classification of named entities in crime domains can give some important information about the crime, and such information can help to facilitate crime investigation. It also provides basic information for crime analysis. In addition, these entities can be used by other NLP applications in a crime field such as text summarization, relationship extraction, Information Extraction (IE), Information Retrieval (IR) and Question Answering (QA) which can give a more significant analysis for the crime.

Named Entities (NEs) are every proper noun existing in documents. NER is an important task in NLP areas, especially for information extraction. It is a system which identifies and classifies vocabularies or sequences of words indicating a conception of entity, such as persons' names, organization names, location names, dates and times. The NER task in a specific language is regularly achieved by the collection of information about the language. For example, in the English language, such knowledge may involve known titles, capitalization of proper names, suffixes or common prefixes, recognition of noun phrases in documents and Part Of Speech (POS) tagging. Techniques that are constructed for a specific language may not be suitable for another language.

Many research investigated NER problem in a variety of languages and domains. However, only a few limited researches have focused on NER for crime text. In addition, when moving to a new domain, the lexical resources should be customized ,the system needs to be modified [1] and the domain-specific features needs to be utilized. According to our knowledge, all of Arabic NER systems are for general domain and there has been no research about Arabic named entity in crime documents. Named entities in a specific domain mean the terms or the phrases that point to concepts relevant to one exact field. For example, protein and gene names are named entities which are of interest to the biomedical field. Another example is the names of chemical compounds which are of interest

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

to the chemistry domain. NER results that get a high rank of accuracy in some field or language may achieve much weaker results in a more diverse context.

In this paper, we attempt to solve the problem of Arabic named entities identification and classification in crime domain automatically by using rule-based NER approach (linguistic approach). This research focuses on inducing and then formalizing a set of syntactical rules and patterns from a small crime corpus for extracting and classifying Arabic named entities crime text. This NER approach uses morphological and contextual evidence, intrinsic indicators and crime-specific knowledge (terms) to create such rules. Evaluation on Arabic corpus of crime news shows that that the method is effective and the performance of the achieved system is satisfactory.

The remainder of this article is organized as follows. Section 2 presents previous work in this field. Section 3 describes the structure of the NER system and introduces the rules. The results of its application in a corpus of crime news are discussed in Section 4 and Section 5. The paper concludes and presents our future plans in Section 6.

2. RELATED WORK

In this section we illustrated some of the pervious works that conducted to Arabic language and crime domain. To the best of our knowledge, most of NER Arabic researches were applied in the general domain and there is no NER Arabic work for crime texts.

2.1. Related Work in the Crime Domain

There are few works which have been published in the crime domain and which had used NLP tools. Chau et al. [2] developed a NER system based on the rules, machine learning (a neural network), statistical-based and lexical lookup approaches. This system aims to identify four types named entities from English police narrative. The system obtained 74% of precision for person type, 59% of precision for address type, 85% for narcotic drug and 45% for Personal property.

Hao Ku et al. [3] introduced an online reporting system that consists of the combination of natural language processing and the epistemic interview approach to get more information from victims and witnesses. This system is developed depending on their own gazetteer lists and JAPE (Java Annotations Pattern Engine) rules which is specific format of GATE (General Architecture for Text Engineering) to describe regular expressions for annotations required for pattern matching . The system can be used by people online to report crime anonymously in English language. Then, the reports are used to offer a purposeful summary for police interrogators to solve crimes.

Pinheiro et al. [4] describe an Information Extraction system on the web, depending on NLP, and it aims to explore the available information about crimes. This system used the Semantic Inferential Model (SIM) that aims to construct a NLP system with an additional layer for understanding texts, which presents an analysis on the explicit and implied content of the text. Furthermore, the system used the tool called WikiCrimes to extract types of crime and the crime scenes from crime news articles existing on the web.

2.2. Related Arabic Work

Many researchers have studied the problem of Named Entity Recognition (NER) in several languages. However, only a few limited researches have focused on NER in Arabic documents due to limited amount of progress made in Arabic natural language processing in general, and the lack of resources for Arabic named entities.

ANERsys is a NER system that has been built totally for Arabic texts based on maximum entropy and n-grams. In addition, the system relied on the gazetteers and the exact Arabic language dependent heuristics. The system is trained and evaluated using the personal gazetteers (ANERgazet) and the personal training and testing corpora (ANERcorp). The researchers obtained the baseline results: 51.39% of precision, 37.51% of recall and 43.36% of f-measure. However, when they used ANERsys (without using ANERgazet) on the ANERcorp test, they achieved a precision result of 62.72%, recall of 47.58% and f-measure of 54.11%. Unlike when they applied ANERsys (using ANERgazet) on the ANERcorp test, they achieved a precision of 63.21%, recall of 49.04% and f-measure of 55.23% [5].

Mesfar [6] developed an Arabic NER system that consists of a syntactic parser and morphological parser that are constructed within the NooJ linguistic development environment. The environment included huge dictionaries, grammars, and parses corpora in real time. The system is used to classify numerics, dates, known proper names and unknown proper names in standard Arabic text. An evaluation process was applied on part of their corpora that was collected from the newspaper "Le Monde Diplomatique" in Arabic version. It showed the following scores for Person names: precision 92% recall 79% and the F-measure 85%.

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

Benajiba at el. [7] described a NER system using Support Vector Machines (SVMs) and the combination of language independent and language dependent features for an Arabic NER. They measured the impact of the different features independently and in a joint combination across different standard data sets and different sorts.

NERA is an Arabic NER system in the general field. It recognized and extracted 10 named entities in Arabic texts: the location, person name, price, company, date, time, phone number, measurement, file name and ISBN. The system consists of a set of rules that are built by using regular expressions and a dictionary of names that is called the whitelist. The personal corpora are tagged in a semi-automated mode and used to evaluate NERA [8].

Elsebai at el. [9] described the implementation and development of a person name named entity recognition system for the Arabic Language. The system adopted rule-based approach that is dependent on the output produced by the Buckwalter Arabic Morphological Analyser (BAMA). The system also uses a set of keywords that is a guide to the probable phrases that may contain person names. The system obtained 89% of F-measure, 86% of recall, and 93% of precision.

AbdelRahman at el. [10] proposed an Arabic NER system which is built based on the combination of two automatic learning techniques which are a Conditional Random Fields (CRF) recognizer as a supervised method and bootstrapping semi-supervised pattern identification. This system is used to recognize the location, organization, person name, job and other classes.

3. THE NER SYSTEM

Our NER system involves modules for linguistic preprocessing, named entity identification and classification. Figure 1 show the detailed architecture of our system.

3.1. Pre-Processing Modules

The system includes three pre-processing modules that need to be used before the NER task. The utilization of these modules depends on the nature of the input. These modules are used when the input is raw text. The modules are sentence splitting, tokenization, and POS tagging.

The sentence splitting module: in this step the input text is segmented into several sentences. Besides, the boundaries of the sentence can be classified by symbols such as, end of line, punctuation and full stop. As a result of this segmentation the output will be annotations for each boundary as well as annotations for each sentence.

The tokenization module: the tokenization is the procedure of analyzing and splitting the input text into a number of tokens such as, number, word, space, symbol, etc. The Arabic token can be a number as a sequence of digits or a word as a sequence of connected letters or a conjunction or a symbol represented as, ?,}, etc.

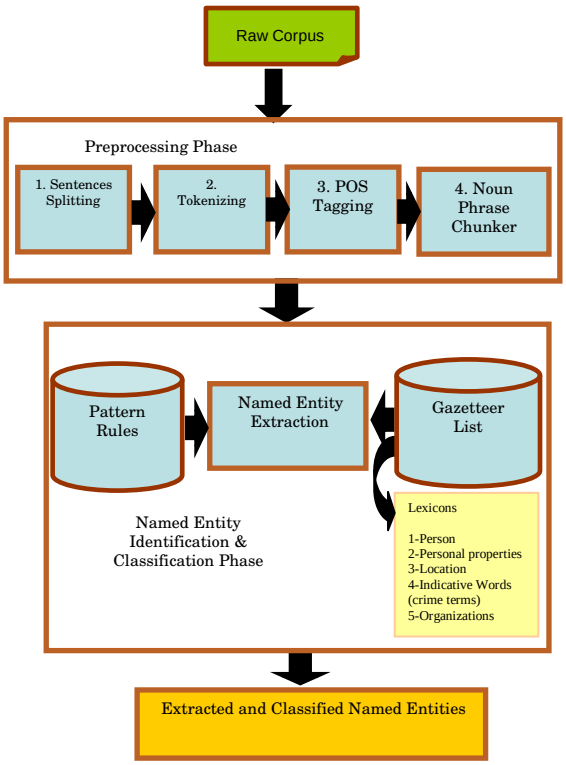

Figure1.The architecture of the system

Part of speech (POS) tagging: For part of speech tagging, a supervised statistical Arabic POS tagger is used[11]. In addition, the morphological features (gender, number, tense) also assigned to each word. This tagger has been trained on our POS annotated corpus which consists of 95 Arabic crimes with size of 19800 words.

Noun Phrase Chunker: The chunker uses the POS tags from the previous components to mark noun phrases.

3.2. Named Entity Identification and classification Module

The identification of named-entities involves the detection of their boundaries, i.e. the start and the end of all the possible spans of tokens that are likely

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

to belong to a named entity. Once the possible named entities have been identified, classification begins. Named entities identification and Classification uses a set of grammatical rules and patterns and gazetteer.

Gazetteer: the gazetteer involves a set of lists contain specific information such as people's names, locations names, organizations names and days of the week. These lists can help the NER system for direct recognition. We used four lists of names (person names, location names, organization names and time).

In addition, the gazetteer also includes five types of lists of indicative verbs and words for named entities that always processed or followed them in text. Because of the nature of our system regarding crime texts, the indicative words and verbs are chosen in both the general and in the crime domains. These lists includes Indicative Verbs for Person names (IVP), Indicative Words for Person names (IWP), Indicative Words for Locations (IWL), Indicative Words for Organizations (IWO), and Indicative Words for Time (IWT). The lists are considered as keywords that can help to identify some entities within documents being those that may come before the entities in the Arabic text.

3.3. Implementation of Rules

We have developed a set of rules specific to the recognition and classification of Named Entities in Arabic crime documents. The rules are induced and formalized by analyzing data from 65 Arabic crime articles comprising 13300 words. These rules are applied using the developing corpus. The input data is prepared in a specific format and each line contains only a POS tag matching with the word in the sentence. The rules are applied to the input text parts. Each rule is applied if its conditions are met for recognition of the NEs. The rules are mainly dependent on three terms which are:

- The POS tag for each word in the input text.
- The indicative verb and word lists such as the where the development of these lists play a central role in the construction of rules where they are used as keywords to find the positions of NEs in the text.
- The lists of names (Gazetteer) which are lists of known person names, location names, organization names and time names that are used for direct recognition.

The rules have been written in regular expression formulas. The rules have described some sequences of tagged words to identify the five types of NEs which are covered by our rules which are person name, location, organization, date and time.

4. EVALUATION

To properly evaluate the rule-based named entity recognition for Arabic crime texts, we have developed a small corpus by choosing a set of Arabic crime documents from four Arabic newspapers (Albyan, Aljazeera, Okad and Gorena) that currently existing on the net.

The POS annotated corpus consists of 95 Arabic crimes. 65 crimes have been used for training and 30 for testing. We have manually labelled the whole of the person names, locations, organizations, dates and times that exist in the training corpus. Then, we have analyzed the training corpus and developed a number of rules to recognize the needed named entities automatically.

The standard evaluation measures in the information extraction, F-measures, precision, and recall, are used to evaluate the proposed model. Recall is defined as the ratio of number of named entities words retrieved and classified to the total number of named entities words actually present in the test corpus (gold standard). Precision is the ratio of number of correctly retrieved and classified named entities words to the total number of named entities words retrieved by the system.

These two measures of performance are combined to form one measure of performance, the F-measure, which is computed by the weighted harmonic mean of precision and recall.

$$F_{\beta} = \frac{(\beta^2 + 1) R P}{\beta^2 R + P}$$

5. RESULTS AND DISCUSSION

The rule-based NER system for crime documents was applied on the testing set which consists of 30 crimes (6500 words). Table 1 shows the accuracy of our system in terms of the precision, recall, and F-measure for each class of the named entities (person name, location, organization, date, time) in the Arabic language.

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

Table 1 Performance of the Proposed System

Class	Precision	Recall	F_1
PER	96	88	91.83
LOC	100	96	97.96
ORG	90	92	90.99
Date	100	88	93.62
Time	67	80	72.93
Overall	91	89	89.46

From Table 1, it can be shown that the result of our system is satisfactory when compared against the human evaluation where the recall is 96% in the location class, 92% in the organization class, 80% in the time class, 88% in the person name and date classes. The F-measure is 92% in the person name class, 98% in the location class, 91% in the organization class, 94% in the date class and 73% in the time class. Lastly, the precision of our system is 96% in the person name class, 90% in the organization class, 100% in the location and date classes.

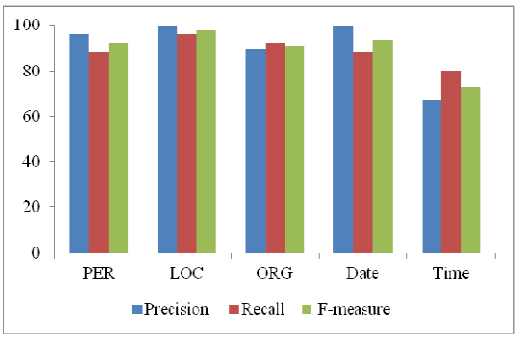

Figure 2. Performance (Precision, Recall, F_ measure) Of the Proposed System for Each Class

The best overall results (Fig.1) have been achieved for location names, and date names, 0.97 and 0.93 respectively, whereas results for time names are the lowest ones.

In brief, the overall performance of the system in all classes of NE is around 89.4% of F- measure. However, the other measures were about 91% for precision and 89% of recall. The overall performance (F-measures) for these types of named entities is quite satisfactory compared to relevant work for crime texts [2].

6. CONCLUSION

Our research focus is on building an Arabic NER of the crime field. Recognition and classification of named entities in this domain provides basic information for crime analysis and can be used by other NLP applications which can give a more significant analysis for the crime. This paper contributes towards the design and implementation of a rule-based NER system to extract and classify NEs from Arabic crime documents. We have designed a set of syntactical rules and patterns by considering features such as prefix and suffix current word, morphological and POS information, information about the surrounding words and their tags and also by utilizing predefined crime and general indicator lists and an Arabic named entity annotation corpus from crime domain. Our next step is to integrate the rule-based NER system with machine learning techniques and to embed it within a crime analysis system.

REFRENCES:

[1] Farmakiotou, D. Et Al.,2000, "Rule-Based Named Entity Recognition For Greek Financial Texts," 2000, Proc. of the Workshop on Computational lexicography and Multimedia Dictionaries. 2000,PP. 75-78.

[2] Chau, M., Xu J. J., & Chen, H. 2002. Extracting Meaningful Entities from Police arrative Reports, in ACM International Conference Proceeding Series. Los Angeles, California: Digital Government Research Center (129) 1-5.

[3] Hao Ku, C., Iriberri, A. & Leroy Chih, L. 2008. Crime Information Extraction from Police and Witness Narrative Reports, 2008 IEEE International Conference on Technologies for Homeland Security.

[4] Pinheiro, V., Furtado, V., Pequeno, T. & Nogueira, D. 2010. Natural Language Processing Based on Semantic Inferentialism for Extracting Crime Information from Text. 2010 IEEE.

[5] Benajiba, Y., Rosso, P. & Bened´ı Ruiz, J. 2007. ANERsys: An Arabic Named Entity Recognition, System Based on Maximum ntropy , Springer-Verlag Berlin Heidelberg: 143–153.

[6] Mesfar, S. 2007. Named Entity Recognition for Arabic using syntactic grammars, Springer-Verlag Berlin Heidelberg, Paris, France 305-316.

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

[7] Benajiba, Y. 2009. Arabic Named Entity Recognition. Ph.D. thesis, Universidad Politecnica de Valencia 1-206.

[8] Shaalan, K. & Raza, H. 2008. Arabic Named Entity Recognition from Diverse Text Types. Springer-Verlag Berlin Heidelberg .

[9] Elsebai, A. 2009. A Rules Based System for Named Entity Recognition in Modern Standard Arabic. Ph.D. thesis, University of Salford, UK.

[10] AbdelRahman, S., Elarnaoty, M., Magdy, M., & Fahmy, A . 2010. Integrated Machine Learning Techniques for Arabic Named Entity Recognition. IJCSI International Journal of Computer Science Issues.7(3): 27-36.

[11] Albared, M., N. Omar, And M.J. Ab Aziz, Improving Arabic Part-Of-Speech Tagging Through Morphological Analysis, In Proceedings Of The Third International Conference On Intelligent Information And Database Systems - Volume Part I. 2011, Springer-Verlag: Daegu, Korea. P. 317-326.