Information Extraction from Arabic News

Hala Elsayed[1], Tarek Elghazaly[2]

[1] Computer and Information Sciences Dept., ISSR, Cairo University
Cairo, Egypt

[2] Computer and Information Sciences Dept., ISSR, Cairo University
Cairo, Egypt

Abstract

Information Extraction (IE) is concerned with finding of specific facts from collections of vast unstructured texts found in the web and in large documents. The Named Entity Recognition (NER) is a sub-problem of the Information Extraction (IE). The recent research in information extraction are growing and also there are interests in the Named Entity Recognition (NER) which helps in extracting the desired information from massive texts and hence extracting entities is an important task in the Natural Language Processing (NLP).

The Arabic Language needs to perform more researches in information extraction domain and hence we introduce this research. The experiment is concerned with extraction entities and entities relation extraction from the Arabic text. We used in our experiment text from Arabic news in Egyptian Arabic newswire. The paper introduced a method for extracting numerous unknowns using entity and entities relation from Arabic Corpus that is generated from Egyptian Arabic newswire to extract Information using the Named Entities and Entities Relation in Arabic language. The experiment contained nearly 625368 entries; the number of sentences was 36423 and the selecting sample was about 3400 sentences representing the crimes news. In the results we obtained some information that is considered a tool for a decision-maker in analyzing the text.

Keywords: *Information Extraction (IE), Natural Language Processing (NLP), Named Entities Recognition (NER), Corpus, Gazetteers.*

1. Introduction

The World Wide Web (WWW) contains Arabic texts and Arabic news for over than 300 million people who exchange information and knowledge which reflect the real-world, wherever the texts news include entities and un-known hidden relations between entities [1][2][3][4]. Recognizing the relations between entities is among the important tasks on the Web such as Information Retrieval (IR), Information Extraction (IE).

Information Extraction (IE) is the process of identifying and recognizing within text instances of special classes of entities and of predictions involving these entities [4][5].

One main objective of the Information Extraction (IE) is helping users to rapidly identify relations from massive texts. These relations can be represented in pair entities or set of entities. The Message Understanding Conferences MUC-7, define the following three types of relations between pair of entities: Person-employer, maker-product, and organization-location [8].

The Arabic Named Entities Recognition suffers in tasks as collecting huge Arabic corpora, gazetteers and so on. Therefore, Arabic named entities recognition researches try continuously to develop and improve named entities recognition in the Arabic language [6].

The Named Entity Recognition (NER) is the task of locating and classifying names in text [10], and is concerned with finding relations between entities [7][8]. In this research we assumed that the entity extraction and the entities relation extraction are not enough for extracting minimum information of massive un-structured Arabic text.

2. Motivation

Currently, the Arab World has many resources as newspapers, electronic newspapers, radio, television and satellite for exchanging information. But unfortunately news may be interpreted in an unfair way without analysis using statistics and for that reason we introduce this paper concerning how to extract information automatically from Arabic text news through the use of the Named Entry Recognition (NER).

3. Named Entities

The Named Entities is a sub-task of Information Extraction (IE) where previous researches in this field has focused on seeking for entities, entities as proper names, locations, organizations, vehicles, books, cats, biomedical entities as gene, organisms, malignancies, chemicals, expressions of emails, time, quantities, monetary values, percentages, measure, abbreviation and relations among entities as person name and organization.

A Named Entity is a piece of text string which refers to an Entity [8][9]. Let $N\mathcal{E}$ be the set of Named Entities in the text; $N\mathcal{E}_t$ be the set of all named entities of type t; Let T be a bag of named-entity types; let $\mathcal{R}$ be the set of all named entities relation where:

$$N\mathcal{E} = N\mathcal{E}_{t1} \cup N\mathcal{E}_{t2} \ldots\ldots \cup N\mathcal{E}_{tn}$$

$$\mathcal{R} \subseteq \prod_{t \in T} N\mathcal{E}_t$$

This paper represents the types of NER and locations and drugs and the relation between both locations and drugs.

$$\mathcal{R} \subseteq N\mathcal{E}_{location} \times N\mathcal{E}_{Drugs}$$

4. Building Corpus

The corpus is consisted of set news of Al-youm7 journal that is for our experiment. The following steps were utilized for constructing: Corpus, data resources, pre-processing Arabic news.

4.1. Data Resources

Constructing the corpus was from the Arabic news where the text was collected using a crawler program. In the research, we chose the newswire web pages of Al-youm7 in the address http://www.youm7.com, September 2012.

4.2 Pre-Processing Data

We selected the incidents news to experiment so that we picked the concerning news with incidents news. For obtaining pure text to build Arabic corpus which passed with the following four phases: The first phase is removing the unwanted texts, images and signs... and so on. The second phase is separating and splitting the text into heterogeneous segments to be easily-handled and helping for processing Arabic text. Table (1) shows the division and organization the data into the following four portions where the first segment expresses the index of news; the second segment is the headline news; the third segment is the date-time of news and the fourth segment includes the news details.

Table 1: Example of News Organized into Heterogeneous Portions

ID	Headline News	Date-Time of news	Details of the news
1	ضبط كميات كبيرة من البويات منتهية الصلاحية قبل بيعها للجمهور بطنطا.	الجمعة، 28 سبتمبر 2012 - 18:43	تمكن ضباط مباحث التموين بالغربية من ضبط كميات كبيرة من "البويات" المنتهية الصلاحية، داخل محل لبيع تركيبات دهانات السيارات بالكمبيوتر، بقصد إعادة خلطها وتركيبها وبيعها للجمهور.
2	ضبط 1392 مخالفة متنوعة فى حملة مرورية بالمحلة وطنطا.	الجمعة، 28 سبتمبر 2012 - 18:36	شنت مديرية أمن الغربية حملة مرورية مكبرة لضبط المخالفات المرورية والسير بدون لوحات وعكس الاتجاه بمدينتى المحلة وطنطا.
3	حبس سائق 4 ايام بعد قيامه بطعن زوجته داخل كنيسة بالمحلة.	الجمعة، 28 سبتمبر 2012 - 18:18	قرر إبراهيم الجمل، وكيل نيابة ثان المحلة، حبس السائق المتهم بطعن زوجته داخل كنيسة الأورام 4 أيام احتياطيا على ذمة التحقيقات، وسرعة تحريات المباحث حول الواقعة.

5. The Experiment

In our experiment, we are interested to extract Named Entities (NE) types of location entity type and drugs entity type but on other hand we are interested in extracting the relation among location entity type and drugs entity type which serves the idea of Information Extraction (IE). During the experiment, headline news contained small and complete shortcut news, and we decided to add semantic annotation to clear the types of entities and therefore we annotated the Arabic text such as location, drugs and weight.., and so on. Table (2) contains examples of Tag Set.

Table2: Examples of Named Entities Tagset

NE Tag	Meaning	Example
<LOC>	Location Entity	<LOC>اسوان <LOC >بالمحلة
<DRG>	Drugs Entity	<DRG>مخدرات
		<DRG> ترامادول
<WGH>	Weight Entity	<WGH>كيلو <WGH>طن

Extraction Entity: We added a tag next to every different locations, drugs, and weights for extract entries.

Extraction Relation: The system can extract two or three entities which appeared regardless whether there are relations between entities or not.

6. Experimental Results

The system extracted entities of location, drugs, weight and indicated whether there are relations between entities in the news text. Table (3) shows examples of tag set in the Arabic text.

We calculated the frequency named entities and entity relations that were extracted from the Arabic news text. The system is the extracted locations entities. We obtained around 58 different locations entities as shown in Table (2) so that set threshold 3% to accept the location entity, with applied the threshold we obtained the results of 12 different locations as shown in Figure (1).

Table 2: The Locations Entities Tagset Extraction from the Crime News

ID	Location Entity	Frequency (%)
1	واحة سيوه<Loc>	0.18%
2	وابوقتادة<Loc>	0.18%
3	كفرالشيخ<Loc>	3.49%
4	قنا<Loc>	2.94%
5	طنطا<Loc>	2.02%
6	شرم الشيخ<Loc>	0.18%
7	سوهاج<Loc>	3.13%
8	رفح<Loc>	0.74%
9	خليج نعمة <Loc>	0.74%
10	بينها<Loc>	0.18%
11	بورسعيد<Loc>	1.29%
12	بمطروح<Loc>	0.37%
13	بمدينةنصر<Loc>	0.92%
14	بلبيس<Loc>	0.55%
15	بشبين<Loc>	0.18%
16	بسيناء<Loc>	0.92%
17	بسمنود<Loc>	0.74%
18	براس<Loc>	0.37%
19	بدسوق<Loc>	0.18%
20	بحلوان<Loc>	1.10%
21	بحدائق<Loc>	0.37%
22	ببولاق<Loc>	0.55%
23	ببنى سويف<Loc>	1.47%
24	ببسيون<Loc>	0.55%
25	بالوادى<Loc>	1.10%
26	بالمنصورة<Loc>	1.10%
27	المنيا<Loc>	4.04%
28	بالمطرية<Loc>	0.37%
29	بالمحلة<Loc>	2.39%
30	بالقليوبية<Loc>	2.39%
31	بالفيوم<Loc>	1.29%
32	بالغربية<Loc>	3.13%
33	بالعياط<Loc>	0.55%
34	بالشيخ<Loc>	0.18%
35	بالخانكة<Loc>	0.55%
36	بالجيزة<Loc>	0.18%
37	بالتبين<Loc>	0.18%
38	بالبساتين<Loc>	0.37%
39	بالاقصر<Loc>	1.65%
40	بالاسماعيلية<Loc>	0.18%
41	باسيوط<Loc>	5.51%
42	باسوان<Loc>	2.39%
43	بابوالنمرس<Loc>	0.18%
44	المنزلة<Loc>	0.37%
45	المشابك<Loc>	0.18%
46	المرج<Loc>	0.18%
47	القاهرة<Loc>	5.15%
48	العريش<Loc>	1.84%
49	العاشر<Loc>	2.02%
50	الشرقية<Loc>	10.11%
51	السويس<Loc>	3.31%
52	الدقهلية<Loc>	0.92%
53	الحوامدية<Loc>	0.55%
54	الجيزة<Loc>	3.49%
55	البدرشين<Loc>	1.47%
56	البحيرة<Loc>	4.23%
57	الاسماعيلية<Loc>	7.72%
58	الاسكندرية<Loc>	5.70%

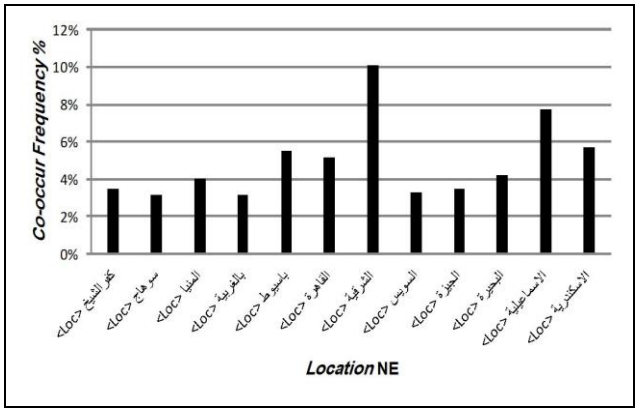

Figure 1: Information Extraction for Locations Threshold

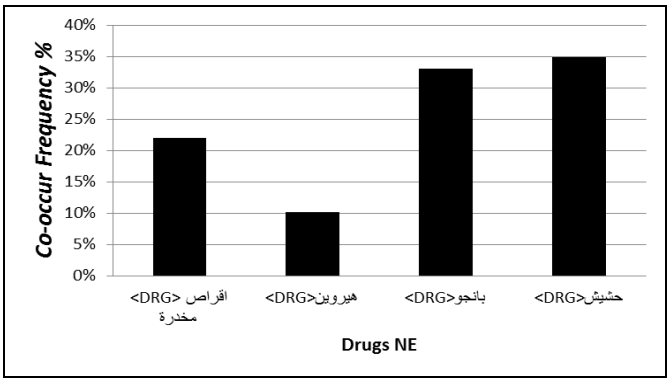

Figure 2: Drugs NE

6.1. Relation Extraction

Relation Extraction means finding relations among entities in text, the idea of entities relations extraction introduced in MUC-7, 1998[11] for example the relation extraction person and company. The relations have been studied starting from Year 2002 covering the ACE evaluations.
In our experiment, we obtained the results of the relation between drugs and location as shown in Table (3).

Table 3: The Relation Extraction locations Entities and Drug Entities

Location	Drugs			
	اقرص مخدر<DRG>	حشيش<DRG>	بانجو<DRG>	هيروين<DRG>
الاسماعيلية<Loc>	13.04%	10.53%	4.55%	16.67%
بسمنود<Loc>	4.35%	-	9.09%	-
كفرالشيخ<Loc>	-	-	27.27%	-
البدرشين<Loc>	-	-	4.55%	16.67%
القاهرة<Loc>	21.74%	5.26%	4.55%	16.67%
الشرقية<Loc>	-	15.79%	18.18%	16.67%
سوهاج<Loc>	-	10.53%	4.55%	-
بالغربية<Loc>	4.35%	5.26%	13.64%	16.67%
المنيا<Loc>	-	-	4.55%	-
ببسيون<Loc>	-	-	9.09%	-
بمطروح<Loc>	-	5.26%	-	-
البحيرة<Loc>	-	10.53%	-	-
السويس<Loc>	-	5.26%	-	-
باسوان<Loc>	4.35%	15.79%	-	-
بالاقصر<Loc>	-	5.26%	-	-
اسيوط<Loc>	13.04%	5.26%	-	-
الاسكندرية<Loc>	4.35%	5.26%	-	-
المنيا<Loc>	8.70%	-	-	-
بالوادى الجديد<Loc>	4.35%	-	-	-
بالقليوبية<Loc>	4.35%	-	-	16.67%
الجيزة<Loc>	4.35%	-	-	-
العريش<Loc>	4.35%	-	-	-
بالمنصورة<Loc>	4.35%	-	-	-
قنا<Loc>	4.35%	-	-	-
Total	100.00%	100.00%	100.00%	100.00%

7. Summary

Building a system for extracting entities for helping in Information Extraction (IE) which is summarized into the following two steps: The First Step deals with news by annotated text which indicated the entities targets. The Second Step is to split the text into segments on news sentences as shown in the next context news. In each sentence is split words into candidate instances.
If there are candidate instances of intended entities so the system is extracted entities. For extracting the relations, we decided to extract entities as location and drugs or more and therefore we added a tag set into the text to extract other entities from the text.
In our experiment, the system is searching for two or more candidate instances, if the candidate instances are intended entities so the system is extracted entities as relation entities. See Figure (3).

$\text{ContextNews}_{1\text{-}n} = S_1, S_2, S_3 \ldots\ldots\ldots, S_n$, where S_1 is first sentence, S_n is the last sentence.

$\text{ContextSentence}_{1\text{-}n} = W_1, W_2, W_3, \ldots\ldots\ldots, W_n$, where W_1 is the first word in sentence, W_n is the last word in sentence.

Filtering the entity type which have the desired Tag set and extracting it from each sentence; and filtering the entities relation extract the whole entities types which have the desired Tag set from each sentence.

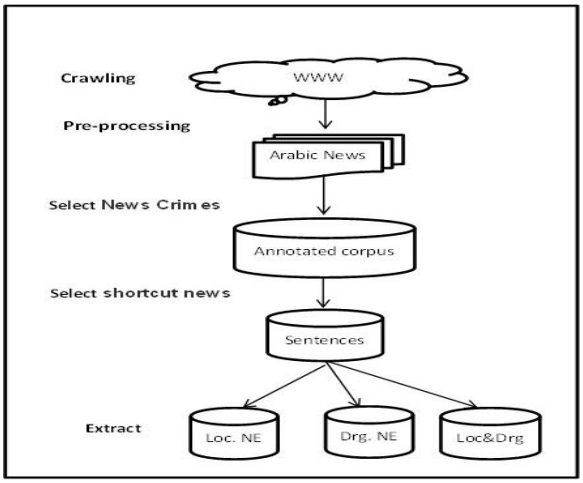

Figure 3: Steps of Extracting NE, NE Relation

8. Conclusions

In this paper, we presented the Information Extracting through using Named Entity Recognition (NER). The extraction of information depends on annotating the target entities which supports the extracted hidden information and extract information automatically from the massive Arabic text which helps the decision-maker.

References

[1]A. Mansouri, L. S. Affendey, and A. Mamat, "Named Entity Recognition Approaches," vol. 8, no. 2, pp. 339–344, 2008.
[2]S. Abdelrahman, M. Elarnaoty, M. Magdy, and A. Fahmy, "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," vol. 7, no. 4, 2010.
[3]M. Ipalakova, "A dissertation submitted to the University of Manchester for the," pp. 1–89, 2010.
[4]J. Tang, M. Hong, D. Zhang, B. Liang, and J. Li, "Information Extraction : Methodologies and Applications Keyword List."
[5]M. Ipalakova, "A dissertation submitted to the University of Manchester for the," pp. 1–89, 2010.
[6]S. Abdelrahman, M. Elarnaoty, M. Magdy, and A. Fahmy, "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," vol. 7, no. 4, 2010.
[7]A. T. Mitchell, "Data Analysis Project : Semi-Supervised Discovery of Named Entities and Relations from the Web Sophie Wang," pp. 1–30, 2009.
[8]J. L. Leidner, G. Sinclair, and B. Webber, "Grounding spatial named entities for information extraction and question answering," 2002.
[9] S. Krause, H. Li, H. Uszkoreit, and F. Xu, "Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web."
[10] D. Downey, M. Broadhead, and O. Etzioni, "Locating Complex Named Entities in Web Text," 1996.
[11]A. Ekbal and S. Bandyopadhyay, "Named Entity Recognition using Support Vector Machine : A Language Independent Approach," pp. 155–170, 2010.