# A Rule-Based Entities Recognition System for Modern Standard Arabic

**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

The Named Entity Recognition (NER) is a task in Information Extraction (IE). The Named entity recognition has become very important for natural language processing. The named entity recognition is defined as the detection and classification of entities from un-structured text where for the Arabic language, the named entity recognition is new in the natural language processing although it has progressed in other languages such as English language. The named entity recognition researchers have become of great interest in recent years for Arabic natural language processing because the named entity recognition plays an essential role for both the information extraction systems and the question answering systems. In this paper, we designed a system which enhanced the named entities recognition for Arabic language where the system was developed for Arabic nouns and entities extractions. The nouns extraction system is based on Arabic morphological which uses no gazetteers where the system is combined with entities extraction system depending on gazetteers. The systems extracts nouns according to morphological Arabic and classify them into: person name entities, title entities, countries entities, cities entities, nationality entities, date and time entities for open text. The system extracts entities in the modern standard Arabic text by two ways: the first way is through using classifying entities annotation in the text; and the second way is through adding entities tag set in the text. The system achieves results in an average recall of 84%.

***Keywords:*** *Message Understanding Conference (MUC), Gazetteers, Named Entities Recognition, Corpus.*

## 1. Introduction

Modern Standard Arabic is the language widely used across the Middle East, North Africa, Horn of Africa and it is one of the official fifth languages used in the United Nations. There are more than 300 million people who speak Arabic all over the world. In the Arab world, there are increasing numbers of publishers of books, website, newspapers, magazines and official documents of Arabic language. It is difficult to extract information from the Arabic text and therefore information extraction is of great importance which leads to the mission of named entities recognition which is of increasing importance.

Arabic Language is one of the Semitic language families [10]. The start of the classical Arabic era is usually calculated from the sixth century which saw a vigorous flourishing of the Arabic literary. In the seventh century, Prophet Muhammad came with the revelation of verses which constituted the holy book (Quran) which was considered as the important book of the classical Arabic. The modern period of Arabic dates ranged approximately from the end of the eighteenth century whereas modern standard Arabic was developed and became the written norm for all Arab countries as well as the major medium of communication for public speaking. The grammar of both Classical Arabic and Modern Standard Arabic are largely similar in its particulars [9].

Increasing online modern Arabic documents online needed information extraction (IE) which involves automatically identifying selected entity types in free text. The Information Extraction (IE) system used to extract valuable information or knowledge from texts may include the following different types of entities such as person name, title, events and more of the entities [2] [4], and therefore, Named Entities Recognition (NER) is very important to analyze the Arabic text.

The Named Entity is recognition detection and classification of entities such as organizations, persons, places, money, measures, dates and time, expressions, and numbers [1] [5] [7] [9] [11] which was first introduced in the 6th Message Understanding Conference (MUC-6) [6] [8] in November, 1995 and is now widely used and plays a very important role in many areas of Natural Language Processing especially in information retrieval, information extraction, text summarization, text classifications, and question answering systems.

The Paper is organized as follows: Section Two presented the Arabic grammar. Section Three presented the System Architecture and described some of its components. Section Four presented the experimentation.

## 2. Arabic Grammar

Modern Standard Arabic has a grammar which is similar to Classical Arabic, but Classical Arabic has no easy meaning and understanding and has a more complicated form which is found in: Quran, ancient literature, old religious writing, old text and Islamic religious texts. In the new horns, classical Arabic was developed across more than one generation usage of a language developed from classical Arabic and which has converted to modern standard Arabic. This Modern Standard Arabic provides a universal form of the language that can be understood by everybody and is commonly used in: radio and TV news broadcasts, films, plays, poetries, newswires, and conversation between Arabic-speaking people of different dialects.
The Arabic Language is considered as one of the semantic languages where it is has special characteristics but has no capitalization letters and lacks standardization in writing and has free arrangements of word order, and therefore it is a complicated process to extract the named entities from Arabic language. Consequently, the named entity recognition needs more researches along this domain. Also, the Arabic language has a complex morphological system [3] that makes Arabic a very difficult language which contains: prefix, suffix, and affix, so therefore Arabic language is considered as a very strongly structured text.

### 2.1. Gender Nouns

(Nouns) in Arabic language are the names of things, which can be objects, people, or places. Nouns in Arabic, both human and non-human, are either masculine or feminine. Usually, if a singular noun ends in a “ta-marbuta” "ة" then it is expressed as a singular feminine noun. For example the noun of “engineer” is in singular feminine “مهندسة” but is in singular masculine “مهندس”.

### 2.2. The Definite Article

In English, we have definite nouns which are usually preceded by the word 'the' (i.e. the student, the tables, the sun) as well as indefinite nouns (i.e. a student, tables). Arabic Language also has definite nouns and indefinite nouns. The definite nouns are adding an article tool called “alef-laam” “ال” which joins with the token that it precedes. The token is connected with the definite article tool “alef-laam” “ال” that is referring to the nouns.

### 2.3. Plural

In the Arabic Language, Plural is the form which refers to more than two objects or persons. There are two types of plural noun and plural adjective forms: they are either Regular Plurals or Broken Irregular Plurals. The plurals also divided into masculine or feminine where the plural masculine regular ends up with waaw-non “ون” or ya-non“ين” For example, the meaning of the word “scholars” it can be either “عالمون” or “عالمين”; and the regular plurals feminine nouns end up in alef-taa “ات” for example “papers” “ورقات”.

### 2.4. Tanween

The tanween is a hold sign which exists at the end of a noun only and the tanween is not held at the end of the verb. The sign of tanween is the doubling short vowel signs. The tanween is either fat’ḥatayn " ـً" or ḍamm’atayn " ـٌ" or kas’ratain "ـٍ" depending on the Arabic morphology rules.

### 2.5. Characters not connected in the verb

In Arabic Language, there are letters are never connected to the verb contrariwise. There are characters which are connect to a noun such as “le-el” “لل” at the start of the token, and “alef-hamza” "اء" at the end of the token. Otherwise, there are letters connected to the stem verb in prefix for example: ya” “ يـ”, and “ta” “تـ”and suffix for example “waw-non” “ون” and “yaa-non” “ين”. Section Four explains our nouns algorithms in details.

### 2.6. Sentence

In Arabic Language Grammar, there are two basic types of sentences based on the sentence's first word. The first type of sentence is the Nominal Sentence, and the second type is the Verbal Sentence where the nominal sentence starts with a noun, and the verbal sentence starts with a verb.

## 3. System Architecture

Our system is designed with two sides: the First-side was the manipulation of the Arabic nouns extraction that is processed in the Arabic morphology and grammar rules without using any gazetteers. The Second-side was the manipulation of the Arabic noun extraction that is processed using the gazetteers where the rules were applied as shown in Figure (1). The System morphology portion is very useful for nouns and verbs extraction through applying Arabic grammar that does not need any gazetteers or the system recognized nouns entities and verbs. The system is capable to add a tag set beside the nouns or verbs where the system recognized nouns entities, nouns and verbs by annotate entities. The system applied the rule-base to the classified nouns into: person name tag, title tag and country tag, date-time tag where the system generates nationality used Arabic grammar rules using derivation from the country entity.
General Architecture for Text Engineering (GATE):

GATE is available on the site: https://gate.ac.uk/. GATE is a language engineering environment developed at the University of Sheffield, GATE system has provided plugin for Arabic named entity extraction. The plugin contains various gazetteers which are useful for named entity recognition tasks. Gazetteers play a role in the task of entity extraction. We developed the part of GATE gazetteers lists by extended it and we composed GATE gazetteers in our system.

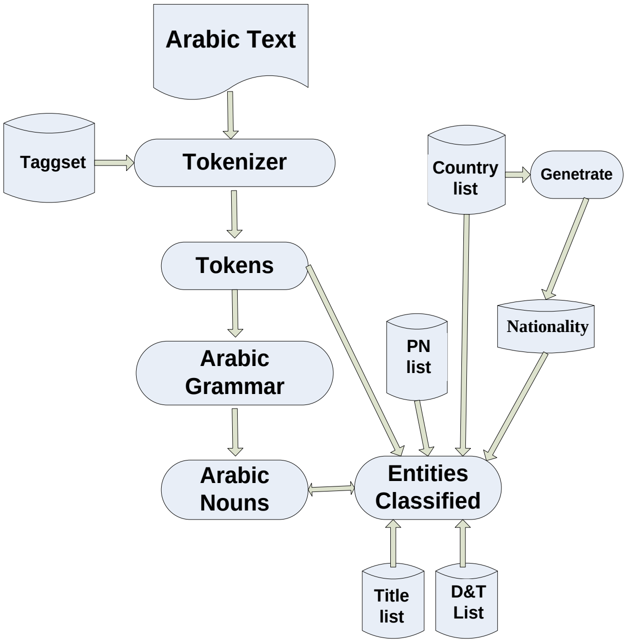

Figure 1 The System Architecture.

# 4. Experimentation

In this section, we will show an experiment and explain the method of extracting nouns by the Next Nouns Algorithms which are used as Arabic grammar rules.

## 4.1. Nouns Algorithms

The system is designed through using the next algorithms to extract Arabic nouns according to the Arabic grammar where the algorithms were applied in Modern Standard Arabic as shown in Table (1).

Table 1: The noun algorithm

| ***The algorithm*** | ***Example*** |
|---|---|
| Read token T from the text<br>IF the token start with “alef-laam” "ال"<br>Then T is noun | T: the north “الشمال”<br>North “شمال” is noun<br>The “ال” is article tool |
| Read token T from the text<br>IF the token start with “ka-alef-laam” “كال"<br>Then T is noun | T: like the water كالماء<br>Water ماء is noun<br>Like” كـ”<br>The “ال” is definite article tool |
| Read token T from the text<br>IF the token start with “waaw-alef-laam” “وال”<br>Then T is noun | T: and music “والموسيقى”<br>Music “موسيقى” is noun<br>The “ال” is definite article tool<br>And “و ” |
| Read token T from the text<br>IF the token end with “alef-ta” “ات”<br>Then T is noun | T: exports ” صادرات”<br>exports “صادرات” is noun |
| Read token T from the text<br>IF the token start with “alef-laam” "بال"<br>Then T is noun | T :in the article “بالمقال”<br>Article”مقال ” is noun<br>The “ال” is definite article tool<br>In ” بـ” |
| Read token T from the text<br>IF the token end with “ta-marbuta -"ة"<br>Then T is noun | T :machine “آلة”<br>machine “آلة” is noun |
| Read token T from the text<br>IF the token with tanween<br>Then T is noun | T: Mohamed “محمدٍ”<br>Mohamed “محمدٍ” is noun |
| Read token T from the text<br>IF the token end with “alef-hamza” "اء" except limited number of verbs such as came “جاء”, will”شاء”, remote”ناء”<br>Then T is noun | T: desert “صحراء”<br>desert “صحراء” is noun |
| Read token T from the text<br>IF the token start with Le-el "لل"<br>Then T is noun | T: for factory “للمصنع”<br>factory “مصنع” is noun<br>For “لل” |

## 4.2. Classification Nouns

The Named Entity is recognizing and classifying the name of person, date, time and so on. In our system, we use: portions of GATE system, the GATE section is Gazetteers, the system built in lists of person names, titles, cities, countries where in these lists we modified them to be more efficient and we added a new list for nationality and the system derived the nationality from the country list as shown in the Table 2.

Table 2: The Entity of Nationality Algorithm

| ***The Algorithm*** | ***Example*** |
|---|---|
| If the country name t end with “yaa” “ي”, or” yaa-marbuta” ية” or “yaa-waw-non” “ يون” or “ yaa-alef-non” “يان” or “ yaa-taa- alef-non” يتان””or “ yaa-taa-non” “ يتن” or “yaa-yaa-non” “يين ” or “alef-taa” “ات”<br>Then t is nationality | • Egyptian masculine “مصري”<br>• Egyptian feminine “مصرية ”<br>• Nominative two Egyptian masculine “مصريان”<br>• Nominative two Egyptian feminine “مصريتان”<br>• Accusative or genitive two Egyptian feminine “مصريتن”<br>• Nominative Egyptians plural “مصريون”<br>• Accusative or genitive plural Egyptians “مصريين ” |

| | |
|---|---|
| | • plural feminine Egyptians “مصريات” |
| If the noun t is not country name but it is look like country name in the most character except the last character ( the noun end with extra character “alef ا” or “marbuta” “ة”)<br>Then remove extra character form country name<br>If the country name end with “ yaa ي”,or” taa-marbuta ة” or “ waaw-non ون ” or “alef-non ان ” or “ yaa-taa-non يتن” or “ yaa-taa-non يتن” or “alef-taa ات”<br>Then t is nationality | • Russia روسيا Russiansروسيون<br>• Russia روسيا Russians روسيات<br>• Russia روسيا two Russians روسيان<br>• Syria سورية Syrian سوري<br>• Syria سورية two Syrians سوريتين<br>• Syria سورية Syrians سوريون |

## 4.3. Results

The system used a combination of Arabic grammar rules and lists of gazetteers. We derived nationalities entities by using a combination of country list and Arabic grammar as shown in Table (2) where the system used grammar which is considered as guidance to recognize other entities. The experiment showed that it extracting Arabic nouns is shared with entity extraction as shown in Figures (2, 3 & 4).

The Modern Standard Arabic text source which was used in our experiment was Essex Arabic Summaries Corpus EASC corpus where EASC is various articles in UTF-8 coding. The system extracted the nouns and classified entities into: cities entities, countries entities, nationalities entities, date-time entities, person names entities and titles entities.

Table (3) presented the average values for: recall, precision and f-measure (where ß=1) the tested EASC corpus was available on site: http://www.lancaster.ac.uk/staff/elhaj/corpora.htm. System extract nouns entities given by the previous nouns algorithms, city entities, country entities, date-time entities, person entities, and nationality entities derived from country entities. These extracted entities were given by modified gazetteers of the GATE system.

غادر محمد البشير [NN] رئيس جمعية [NN] المحاسبين [NN] القانونيين [NN]
الاردنيين [NN] الى [NN] تونس يوم الثلاثاء [NN] الماضي [NN] ليشارك في الندوة
[NN] الدولية [NN] الخامسة [NN] التي [NN] تنظمها الفيدرالية [NN] الدولية [NN]
للخبراء [NN] المحاسبين [NN] الفرنكوفيين [NN] حول الحاكمية [NN] المؤسسية
[NN] الابعاد [NN] الثقافية [NN] والاقتصادية [NN] الثقافية [NN]

Figure 2. Example of extracting nouns using Arabic grammars without gazetteers.

غادر محمد [PN] البشير [PN] رئيس [Title] جمعية المحاسبين القانونيين الاردنيين
[nationality] الى تونس [City] [Country] يوم [D&T] الثلاثاء [D&T] الماضي ليشارك في
الندوة الدولية الخامسة التي تنظمها الفيدرالية الدولية للخبراء المحاسبين الفرنكوفيين حول
الحاكمية المؤسسية الابعاد الثقافية والاقتصادية الثقافية

Figure 3. Example of extracting entities using gazetteers.

غادر محمد [PN] البشير [PN] [NN] رئيس [Title] جمعية [NN] المحاسبين [NN]
القانونيين [NN] الاردنيين [nationality] [NN] الى [NN] تونس [City] [Country] يوم
[D&T] الثلاثاء [D&T] [NN] الماضي [NN] ليشارك في الندوة [NN] الدولية [NN]
الخامسة [NN] التي [NN] تنظمها الفيدرالية [NN] الدولية [NN] للخبراء [NN]
المحاسبين [NN] الفرنكوفيين [NN] حول الحاكمية [NN] المؤسسية [NN] الابعاد
[NN] الثقافية [NN] والاقتصادية [NN] الثقافية [NN]

Figure 4. Example of combination between noun extraction in Arabic grammars without gazetteers & extracting entities using gazetteers.

$$\text{Recall} = \frac{\#\text{ of correct answers given by system}}{\text{total }\#\text{ of possible correct answers in text}} \times 100$$

$$\text{Precision} = \frac{\#\text{of correct answers given by system}}{\#\text{of answers given by system}} \times 100$$

$$\text{F-measure} = \frac{(ß^2+1)PR}{ß^2 P+R} \times 100$$

Table 3: Experimental Results

| ***Entities*** | ***Recall*** | ***Precision*** | ***Fβ=1*** |
|---|---|---|---|
| Nouns | 76% | 80% | 86% |
| City | 80% | 98% | 88% |
| Country | 99% | 50% | 66% |
| Nationality | 75% | 90% | 82% |
| Date-time | 95% | 95% | 95% |
| Person-Name | 83% | 68% | 75% |
| Title | 87% | 86% | 86% |

## 5. Conclusions

Modern Standard Arabic which was derived from the classical Arabic was widely taught in press written and in newswires. Modern standard Arabic is similar to classical Arabic in their grammar rules with classical Arabic, but modern standard Arabic is easier to understand.

Arabic Language is considered as the fifth globally Language in the world.

Our system is implemented using rule-based approach. The system extracts Arabic nouns according to the Arabic grammar rules where the Arabic grammar rules algorithms were discovered nouns in Arabic language that are of no use in any gazetteers. The system extracted and classified entities into person name entities, title entities, countries entities, cities entities, nationality entities and date and time entities where the system used gazetteers. We experimented the system using text in open modern standard Arabic text. The system achieved results as: the average recall rate equals 84%; the average precision rate equals 81%; and the average F-measure rate equals 81%. In our experiment, the system extracted the nouns according to the Arabic Language grammar without gazetteers and extracted the most of nouns in the free text; but when the system extracted the entities using gazetteers, the system extracted the target entities. We noticed that there are nouns extracted through using Arabic grammar without gazetteers shared with entities that were extracted with gazetteers.

## References

[1] Technology, M. Asharef, N. Omar, and M. Albared, “ARABIC NAMED ENTITY RECOGNITION IN CRIME,” Journal of Theoretical and Applied Information Technology ,vol. 44, no. 1, pp. 1–6, 2012.

[2] B. Kouninef and B. Al-johar, “Extracting Entities and Relationships from Arabic Text for Information System,” Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 11, pp. 641–645, 2011.

[3] I. a. Al-Sughaiyer and I. a. Al-Kharashi, “Arabic morphological analysis techniques: A comprehensive survey,” J. Am. Soc. Inf. Sci. Technol., vol. 55, no. 3, pp. 189–213, Feb. 2004.

[4] J. Huang, G. Zweig, and M. Padmanabhan, “Information extraction from voicemail” Proc. 39th Annu. Meet. Assoc. Comput. Linguist. - ACL ’01, pp. 298–305, 2001.

[5] M. Aboaoga, M. Juzaiddin, and A. Aziz, “Arabic Person Names Recognition By Using a Rule Based Approach,” J. Comput. Sci., vol. 9, no. 7, pp. 922–927, Jul. 2013.

[6] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, “Named Entity Recognition: Fallacies, challenges and opportunities”, Comput. Stand. Interfaces, vol. 35, no. 5, pp. 482–489, Sep. 2013.

[7] P. Hiremath and B. R. Shambhavi, “Approaches to Named Entity Recognition in Indian Languages : A Study,” International Journal of Engineering and Advanced Technology (IJEAT), Vol-3, no. 6, pp. 191–194, 2014.

[8] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, “Malay Named Entity Recognition Based on Rule-Based Approach,” Int. J. Mach. Learn. Comput., vol. 4, no. 3, pp. 300–306, Jun. 2014.

[9] KARIN C. RYDING, “Modern Standard Arabic”, Book, Georgetown University, Cambridge University press, available site: http://www.cambridge.org/eg/academic/ subjects/languages-linguistics/arabic-and-middle-eastern-language-and-linguistics/reference-grammar-modern-standard-arabic?format=HB, 2005.

[10] S. Abdelrahman, M. Elarnaoty, M. Magdy, and A. Fahmy, “Integrated Machine Learning Techniques for Arabic Named Entity Recognition,” IJCSI International Journal of Computer Science, vol. 7, no. 4, 2010.

[11] U. K. M. Bangi, “Arabic Named Entity Recognition Using Artificial Neural Network” Naji F. Mohammed and Nazlia Omar, School of Computer Science, Faculty of Information Science and Technology,” vol. 8, no. 8, pp. 1285–1293, 2012.