International Journal on
Advanced Science
Engineering
Information Technology

Vol.7 (2017) No. 3
ISSN: 2088-5334

Arabic Rule-Based Named Entity Recognition Systems: Progress and Challenges

Ramzi Esmail Salah#, Lailatul Qadri binti Zakaria#

#*Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia.*
E-mail: ramzi@siswa.ukm.edu.my, lailatul.qadri@ukm.edu.my

***Abstract*— Rule-based approaches are using human-made rules to extract Named Entities (NEs), it is one of the most famous ways to extract NE as well as Machine Learning. The term Named Entity Recognition (NER) is defined as a task determined to indicate personal names, locations, organizations and many other entities. In Arabic language, Big Data challenges make Arabic NER develops rapidly and extract useful information from texts. The current paper sheds some light on research progress in rule-based via a diagnostic comparison among linguistic resource, entity type, domain, and performance. We also highlight the challenges of the processing Arabic NEs through rule-based systems. It is expected that good performance of NER will be effective to other modern fields like semantic web searching, question answering, machine translation, information retrieval, and abstracting systems.**

***Keywords*— Arabic named entity recognition; modern standard Arabic; classical Arabic; rule-based**

I. Introduction

Rule-based systems are designed to discriminate Named Entity (NE) by applying rules. Many researchers in initial works for NE extraction usually based on rule-based approach [1]. Named Entity Recognition (NER) is a task specified to recognize Named Entities (NEs) in non-structured language texts. Among the common named entities are personal names, locations, the organization [2], [3]. In this paper, we review work progress in Arabic named entity recognition using rule-based systems. We present a comparison of the previous works in terms of the type of datasets, domains, categories, and performance. This paper also highlights the challenges associated with NER in Arabic text and discusses the techniques used in rule-based approaches.

A. *Definition of Named Entity*

The word of term that vividly indicates any object included in a set or a group of other objects that encompass similar traits. More specifically, the expression of 'named entity' refers to the notion that the scope of a particular entity is limited by certain word even if these entities more than one rigid designator which can symbolize a referent. As a matter of fact, rigid designators usually contain proper names. However, this relies on the interesting domain for which the reference word is related to an object in any domain known as Named Entities (NE). For instance, in bioinformatics and molecular biology, the interest entities are gene products and genes. Hence, NER automatically aims at both identifying and classifying such names in text and deals with them as predefined classes. Table 1 shows that a total of five named entities are included in Arabic sentence. Hence, a named entity could be defined as a word or term that evidently determines an object included in a set of other objects. All the objects in the set are of similar traits. As mentioned earlier, the expression 'named entity' indicates that the word named restricts the scope of entities which include one or several rigid designators which indicate a referent.

TABLE I
Example of a Named Entity

1) رمزي سافر الجمعة الماضي من صنعاء ليلتقي بصالح في دبي
2) Ramzi travelled last Friday from Sana'a to meet Saleh in Dubai

Table 2 Identifying and labelling entities by the named entity extractor.

TABLE II
Classify Named Entity (Sample)

The Named Entity	The Type
(رمزي, Ramzi)	Person
(الجمعة, Friday)	Date
(صنعاء, Sana'a)	Location
(صالح, Saleh)	Person
(دبي, Dubai)	Location

B. NER and NLP Applications

The expression of information Extraction is a considered as an umbrella that includes a specific number of certain tasks that have already been recognized by the Message Understanding Conference (MUC). Generally, the essential and basic task for IE or NLP is what the so called named entity recognition (NER). Hence, the term ‘Named Entity’, is now commonly used in NLP. Besides, NER task was initially presented at the 6th MUC Conference (MUC-6). Additionally, NER for various NEs is playing a vital role in Question Answering (QA), Semantic Web (SW), Information Extraction (IE), Machine Translation (MT), and Information Retrieval (IR) [4], etc. Nevertheless, a text can be contained one or more types of names, including lots of names related to specific domains such as Chemicals names, Sports teams, Organization names, Location names or even Person names. Such names are known as Named Entities (NE). They are also called Named Entity Recognition. NER that automatically aims at identifying and classifying those names in a particular text reflecting predefined classes for these names. To clarify this point, name entity recognition comprises two tasks:

- Identification: refers to the region or direction of name entity in a text.
- Classification: regulates the name entity semantics.
- In the MUC-6, the NER consists of three sub-tasks. These subtasks include:
- ENAMEX (Entity Name Expression): the appropriate names including Persons, Locations, and organizations.
- TIMEX (Time Expression): temporal expressions of time and dates.
- NUMEX (i.e., Numeric Expression): generally, refers to numeric expressions related to money and percentage.

C. Type of Arabic Language

Universally, Arabic has been ranked as the 6th language among the major languages worldwide [5]. Moreover, the Arabic language is the official language of 22 countries in the Middle East [6], and it has a direct religious impact on more than 1.6 billion populations in the whole world[7]. In general, there has been little progress in the Natural Language Processing (NLP) of Arabic Language compared to other universally well-known languages which topped by English the language which has been dealt with as Lingua franca. However, only recently considerable works are available, and they are dedicated to Arabic text NLP and NER. As a matter of fact, three forms of the Arabic language exist namely: 1) Classical Arabic (CA); 2) Modern Standard Arabic (MSA); and 3) Colloquial Arabic Dialects [3]. These three forms could be explained in more details as follows:

1) Classical Arabic (CA): For more than 1500 years, this form has been used as the language of Islam. Besides, most of the Arabic historical and religious texts are handwritten in CA. Additionally, extracting Arabic NE from CA has become an important topic when digitized CA materials to be converted from traditional manuscripts. Currently, classical Arabic is sometimes referred to as Quranic Arabic which was originally the dialects of Arab tribes in Arabic peninsula during the medieval era. With regard to the historical speaking, classical Arabic was widely used until the 9th century, and that was during the Abbasid times. After that, the Arab world has experienced great openness to surrounding cultures and languages such as Turks and Persian and that has led to changes and modifications to classical Arabic to its current form today which is known is modern Arabic. Furthermore, Classical Arabic is also called Ancient Arabic. In fact, classical Arabic is the main branch of Semitic languages which was the communication tool during the pre-Islamic age. It’s greatly imaginative yet sophisticated and those who can speak it pride themselves as they can still use it. Moreover, the classical Arabic grammar is complicated and engaging within its text and its vocabulary and is well-structured and highly layered. Some claim that despite its complex structure its beauty is not matched by any other language on this globe. While the classical Arabic is not fully used as the medium of communication, its importance and significance come from the fact that is the foundation of the modern Arabic used today, and its resources are crucial to its mastery. Furthermore, the cornerstones of the Arabic and Islamic culture and heritage are rooted in the Muslims’ holy book (Quran) and other Islamic and historical resources such as Hadith which are all written in classical Arabic.

2) Modern Standard Arabic (MSA): This is considered as the main type of Arabic widely used today as the official language for communications between governments and public or private sectors and individuals. In principle, MSA can be dealt with as the skimmed version of classical Arabic. As its name indicates, MSA is the evolved version of classical Arabic. Moreover, this form of Arabic is the official language of the twenty-two Arab states for oral and written formal correspondence, occasions, and events. Moreover, the United Nations (NU) account MSA as one of the main official languages. Thus, MSA is the most well-known form these days, and it is the most Arabic language form widely used in education, newspapers, and magazines. Most of the studies that have been introduced to analysis Arabic language documents have been evaluated on MSA information resources such as part of speech tagging [8], collocation extraction [9], noun compound extraction [10] and semantic similarity measurement [11]. Most of the Arabic NLP, containing the research projects of NER, focuses on MSA. Besides, the most significant difference between MSA and CA due to vocabulary, containing NEs, and the conventional written Arabic orthography. What makes it more affordable is that MSA does not involve the short vowels inclusion. Additionally, MSA reveals the needs of the up-to-date expression. This significant flexibility does not exist in CA which usually reflects the older styles as prerequisites. For instance, the Arabic NEs in rare old manuscripts and documents which usually refer to jobs, organizations or places, are dissimilar to the corresponding NEs available in modern documents [12].

3) Colloquial Arabic (CA): This form of Arabic is a region-specific dialect which is mostly used for speaking, as well as social media. Many of its words are derived from MSA. CA is the only spoken form and even if most of the words are derived from MSA, it can be described as region-specific and might be very different from one area of the Arab world to another. For instance, "عبدالقادر" (Abd Al-

Kader) vis-à-vis "عبدالجادر" (Abd Al-Gader) or "عبدالآدر" (Abd Al-Aader).

The Arabic language is distinctive when it comes to its two main types related to either oldness such as Classical Arabic (CA) or newness such as modern standard Arabic (MSA). By examining both of the types, one can identify that there is a significant difference when dealing with lexical meanings, style, and grammatical constructions.

D. Key Issues in Arabic NER.

ANER systems encounter some challenges that are associated with the Arabic language. The important challenges are as follows:

1) Arabic Script: Some of the characteristics of Arabic script impose challenges on ANER. Arabic words are written with connected scripts which are not the case for many other languages such as English.

2) Complex Morphology: It is common in Arabic text as various lexical variations can be obtained from different patterns of agglutination. For instance, the processing of the word (أنلزموكموها/ 'anulzimukumuha) results in a whole sentence: ‘shall we compel you to accept it’. Hence, in this situation, we find that the morphology plays an important role for several natural language processing application especially that require understanding the texts [11], [13].

3) Lack of Resource: In general, the Arabic language lacks the resources to test NER systems [2] and just a few corpora that are created by individual researchers. Some of them are available for free to the public [14] while others are available under license agreement [15].

4) Capitalization Issue: Arabic orthography has no capital letters to distinguish initial letters of proper names like other languages, such as English. Thus, the detection of NEs, either expressed in single words or sequence of words, is difficult [16].

5) Auxiliary Vowels: The Arabic language has some diacritics which represent vowels that are used to alter the meaning of a single word. Hence, totally different word’ meanings can be obtained by only changing the diacritics attached to such a word. For example, the word (Noor- نور) may refer to the proper name (Noor-light), or the verb (enlighten- Nooar), or a person female name.

6) Named Entity Inherent Ambiguity: As with other languages, Arabic computational systems also face the mutual ambiguity problem for two or more NEs, and this adds more critical challenges to designing NLP systems for Arabic NEs [12]. For example, some studies reported that there are 21 various analytical results generated by BAMA for the Arabic word (ثمن/ thaman - price) [17].

II. Materials and Methods

Many rule-based approaches are using human-made rules to extract named entities. Generally, these approaches consist of patterns using grammatical rules. They exploit the handcrafted rules for NER task. These approaches are based on grammar rules coming from the linguistic knowledge, and a list of names to detect complex entities precisely. Additionally, rule-based systems are realized in the form of either finite state transducers or regular expressions [18]. The followings are the important rules and techniques that can be used in Arabic rule-based NER systems. The following two sections summarize the common features used in Arabic ML NER systems as well as related works on these systems.

A. List Look-Up Techniques

List lookups technique depends on lists that can be existed in various forms such as gazetteers, whitelist, and dictionaries which can be obtained using some language corpora.

1) Corpus: Corpus can be used in a large collection of annotated texts where NEs are identified with their types. Corpus can be general/specific or cover single domain (e.g. political, economic). Due to considerable research progress in Arabic language, many corpora are now available such as NooJ [18], ACE and Treebank Arabic datasets [19], ANERcorp, [20], etc.

2) Gazetteer: A gazetteer is a list of defined named entities, and it contains a specific list of names for the particular type of NEs class. For example, Malaysia is in location class of a gazetteer while Sami is in person names class. Gazetteers are also called whitelist or dictionaries [19]. Normally, whitelist contains fixed strings of texts which are taken for granted as NEs without the need for any further identifying mechanisms such as applying grammars rules. The entries in whitelist are either a single word or a multi-word expression such as “ محمد حامد الغزالي/ Mohammad hammed algazali”, but in the dictionary (gazetteers) it may contain only a single name, which is related to NE, also may come in other places not related to NE, for Example: (أحلام/ Ahlam) maybe it comes from a person name or as noun which indicates ‘dreams.’

3) Blacklist (Filter): Blacklists or filters are used as a rejection mechanism for words or string of words that are known to be invalid NEs. For example, [وزير الشؤون المالية المديرالعامthe financial affairs minister and the general manager], here the phrase [المدير العامthe general manager] is invalid NE. The main role of the blacklist is to reject such incorrect NEs which could be words or phrases. Moreover, this rendered the NER system more accurate as they can be used to filter out know expressions that can lead to ambiguity which could not be accurately detected using the available rules and dictionaries in the system.

4) Stop-Words: Normally stop-words list contains a group of frequent words that are associated with NEs but are not valid NEs in themselves and cannot be part of NEs such as prepositions in Arabic. For example,[انتخب في الدورة الاولى/ Elected in the first round], the word (في/ fi - in) belongs to stop-words list [20].

5) Trigger Words (Keywords): Trigger words or keywords are frequent words that normally surround the NEs. They are frequent words that normally come before or after NEs. These keywords are used to identify NEs and can take various forms such as verb list or noun list, for example (قال/ say). The trigger words can be used to identify the type of named entities in short phrases.

B. Linguistic Techniques

In Arabic rule-based systems, the linguistic techniques are normally depending on the rules and pattern of Arabic language writing that are used to extract and recognize NEs. The performance of the linguistic-based NER system depends on how good these rules are in identifying various types of NEs. Some of these rules are as follows:

1) Grammar Rules: Grammar rules are the set of the language grammars which are utilized to recognize named entities. Normally, they are the rules for forming well-structured sentences; hence, applying grammar rules in NER requires the knowledge of the language being processed. The grammar rules are used to identify NEs form text by using predefined rules and patterns to find NEs. Due to the complex grammatical structure of Arabic language, especially the classical Arabic, the use of grammar rules is crucial in NEs recognition. For example, in the sentence: (الملك السعودي عبدالله/the Saudi king Abdullah), the used indicating suffix honorific word is "الملك/ the king" and also the second indicating word is " السعودي/Saudi" which refers to nationality, this leads to identifying the word "عبد الله" as the named entity.

2) Heuristics Rules: Complex heuristic rules are more general rules that depend mostly on the type of the overall rule-based approach used on the NER system. For example, they can be used to further enhance the capability of the applied trigger words in the system such as NEs keywords[20].

3) Morphological Rules: Morphological rules are generated based on the fact that words are normally constructed from stems, prefixes, and/or suffixes. Most Arabic words within a sentence normally contain the word's stem and some other suffix and/or prefix. For example, the word يذهب (ya-dahab) is ي+ذهب, the stem is 'ذهب' + prefix 'ي'. While the above-mentioned grammar rules normally operate on the sentence structure in order to identify NEs, on the contrary, morphological rules which deal with the words structure in order to identify NEs by examining the word stem, affixes and/or suffixes. A well-known approach that many researchers use for NER through morphological rules is called Buckwalter Arabic Morphological Analyser (BAMA) [21]. BAMA mainly operates various tables such as a table for collected Arabic stems, prefixes, and suffixes. For example, the token (بمكه) 'Be-Makkah', the morphological rule considers "ب" as a prefix, hence it will remove it from the word, and the remaining is the root which is (مكة/Makkah) which is a location NE.

C. Rule Base Approach.

One of the earliest work in Arabic NER is known as TAGARAB [22]. It incorporates supporting data and a pattern-matching engine with components of a morphological analysis to identify five types of named entity namely: 1) Person; 2) Location; 3) Organization; 4) Time and 4) Number, with a Recall of 80.8%, a precision of 89.5%, and an F-measure of 85% on random datasets from AI-Hayat. The text used by this method is encoded in ISO-8859-6 that is initially passed through the tokenizer. Consequently, the tokenized stream is administered by what is known as the Name Finder Module. The name finder module consists of two units namely: 1) Morphological tokenizer; and 2) named finder. A stem lists together with feature extraction rule are used as inputs to the finite-state scanner while word lists and pattern-action rules are the inputs into NetOwl Turbo Tag™ pattern engine. The final output is an annotated text with appropriate SGML tags for each extracted item.

Mesfar [18] has developed a system for Arabic NER. He used NooJ linguistic platform for building his system. The system contains a gazetteer, tokenizer, triggers as well as morphological analyzer for the purpose of recognizing proper names, dates, and temporal expressions used in Arabic text. The system evaluated on the part of the Arabic version of what is called 'Le Monde Diplomatique.' The results were reported depending on the types of individual NE. They were as follows: 1) Precision; 2) Recall; and 3) F-measure. The results dwindle between 82%, 71%, and 76% for names of Place and 97%, 95%, and 96% for numerical expressions accompanied with time respectively. Additionally, the overall average accuracy of F-measure is 87%. In this approach, the standard Arabic text is input into Nooj tokenizer which outputs text form. Then, the morphological analyzer process inputs from text forms, the lexicon of simple inflected forms in addition to morphological grammars to generated recognized forms connected with linguistic information. The approach involves the use of gazetteers and syntactic grammars to generate recognized named entities.

Shaalan and Raza [23] established a system known as Person Named Entity Recognition for Arabic (PERA). PERA uses linguistic-grammar-based techniques with Whitelist dictionaries to aim at recognizing person name entities in Arabic text with high and significant accuracy. PERA uses three components namely: 1) name lists called gazetteer; 2) regular expressions (grammar) which form the lexicon, and 3) filtration mechanism through the establishment of some grammatical rules that help exclude invalid names. ACE and Treebank corpus were used along with some Internet resources. PERA achieved an 85.5% precision, 89% Recall and 87.5% F-measure.

Shaalan and Raza [19], [24] enhanced their previous research work and developed Named Entity Recognition for Arabic (NERA). NERA is a hand-crafted rule-based system that uses three components namely: dictionaries, grammar rules (regular expressions), and Filter mechanism. The system uses the same method and the same functionality as PERA, however, NERA could support a total of ten types of NEs, such as names of locations, persons and organizations, Price, Date, ISBN, Time, Phone Number, File Names, and Measurements. In the evaluation, they used several resources from ACE, the Web newspapers, the Quran, and Arabic literature, to build their own corpora, also to make it deeper for extracting semantic information. For persons, locations, and organizations, NERA achieved F-measures of 87.7%, 85.9%, and 83.15% respectively. The system has shown above 90%, an average for all MUC named entities .

Traboulsi [25] developed a system by using local grammars to construct a system Arabic NER. The system finds consistent structures of person names that frequently occur in the news text. It uses several sources for its own corpus (arabiCorpus), which was gathered from the archive of newspapers namely: Al-Hayat, Al-Ahram, Al-Watan (issued in Kuwait)) and At-Tajdid (issued in Morocco). The system also uses some corpus from the Holy Quran, Arabic novels such as 1001 Nights in addition to several medieval medical and philosophical sources. To extract address expressions, time and date from letters, Traboulsi developed his corpus. In this study, the adapted method depends on the use of corpus linguistics, methods, and techniques. Consequently, it goes in line with the so-called local grammar formalism, to pinpoint any patterns related to person names in news texts written in Arabic.

Moreover, Al-Shalabi [26] projected an Arabic NER algorithm aiming at retrieving proper nouns in Arabic through the use of lexical triggers. The research focuses on regional patterns of consideration such as the connector of the name. This algorithm recognizes seven NE types including: 1) person names; 2) major cities; 3) locations; 4) countries; 5) organizations; 6) political parties; and, finally, 7) terrorist groups. Nevertheless, the reported research emphases on person NEs only. Additionally, to pre-process the input for erasing the data and remove affixes, heuristic rules are used by the algorithm. Consequently, triggers of internal evidence, such as connectors of person name, are used to identify the NEs. Besides, the system has been assessed through the use of a total of 20 documents which were selected randomly from Al-Raya newspaper. It was observed that the system reached 86.1% of overall precision.

Furthermore, another study found where [20] also investigated grammar rules in Arabic text. It was developed for person name only using a set of keywords. The system has used pattern matching with Morphological Analyser and achieved better performance in an F-measure of 89% over PERA.

Another study by Zaghouani [27] implemented rule-based on RENAR system which was divided into three levels namely: morphological pre-processing, known name identification and, finally, applying local grammar. The aim beyond this division was to identify unknown named entities. Calculating Precision, Recall, and F-measure was the main target of RENAR which reflected overall results of 87.17%, 65.74%, and 74.95%, respectively where ANERsys 1.0, ANERsys 2.0 for the types person, location, and organization when applying ANERcorp dataset. The overall performance of the system, in terms of Precision, Recall, and F-measure, reached 73.39%, 62.13%, and 67.13%, respectively. The method based on three leading processing steps namely: 1) pre-processing (segmentation rules); 2) lookup of full known names; and 3) recognition of unknown names by means of local grammars in addition to a set of dictionaries. Names repeatedly caught (twice, at least) in the form of long-term multilingual news analysis were manually checked. For the purpose of retaining the name, they were stored in a database.

M. Asharef [28] focused on crime domain; they built a small corpus for the crime. The system use rule-based approach for Arabic NER, which was designed a set of syntactical rules and patterns by considering features such as prefix and suffix of the current word, morphological and POS information, information about the surrounding words and their tags. The system also utilizes predefined crime and general indicator lists as well as Arabic named entity annotation corpus obtained from crime domain. The overall performance of the system, in terms of Precision, Recall, and F-measure reached 91%, 89%, and 89.46% respectively.

Aboaoga and Aziz [28] built a system for Arabic to recognize Person names. This system depends on the trigger words that can be used for identifying the person in different domains such as, manager ‘مدير’, president ‘رئيس’, dean ‘عميد’, player ‘لاعب’, referee ‘حكم’, coach ‘مدرب’, supervisor ‘مشرف’, and teacher ‘معلم/مدرس’. Their corpora were collected from the archives including online Arabic newspaper, koora.net, aleqt.net, and Alquds.net. They used sentence splitter and tokenization with gazetteers. The system has been applied to three domains including politic, economic and sport. The sports domain has a better result compared to the other two domains. In this study experiment, the average of F-measures devoted for recognizing person names reached 92.66, 92.04 and 90.43% in the thee-mentioned domains namely sport, economic and politic, respectively.

More recently, Elsayed and Elghazaly [29] developed an NER system that upgrades the NEs recognition for Arabic specifically Arabic nouns. The extraction method based on two-sided approach, namely, Arabic morphology and grammars both with and without gazetteers. The system identifies person name, title, cities, countries, nationality, date and time in MSA with an average F-measure of 84%. The dataset used was Essex Arabic Summaries Corpus (EASC corpus). The method adapted two kinds of rule-based approach: linguistic and list lookup methods. The former is by using Arabic morphology accompanied with grammar rules without involving any gazetteers. The latter is with the use of gazetteers. A part of GATE system is used as the gazetteers which built in lists of titles, person’s names, countries, and cities. The nationalities NE were derived from countries list.

Other works by Oudah and Shaalan [30] proposed methodology for overwhelming the rule-based NER systems’ coverage drawback aiming at improving their performance and allowing for update using automated rule. The presented mechanism develops the ability of recognition decisions carried out by the hybrid NER system for the sake of determining the rule-based component weaknesses and, thus, derive new linguistic rules which tend to enhance the rule base. The empirical results reflect that the enactment of the improved rule-based system (i.e. NERA 2.0) develops the coverage of the formerly misclassified names including a person, location, and organization where percentages of the named entities types were found to be 69.93, 57.09 and 54.28, respectively. Table 3 shows the summary of literature review for Rule-Base system for the Arabic language.

Recently, the knowledge-based approach [13], [31] has been proposed to classify the concepts in the linguistic resource into NEs and linguistic terms. In this approach, Wikipedia is utilized as a semi-structured resource for determining the named entities such as person, organization, location, events, and media. Since each Wikipedia article is related to several categories, these categories can be exploited to recognize the different named entity types. The trigger words have been used to identify the type of the named entities in short phrases such the categories in

Wikipedia. For example, the keywords of an NE person can contain terms such as (أشخاص‘people’) that can be used to form the patterns (e.g. أشخاص_*) that indicates the categories for some mentioned people, such as أشخاص_على_قيد_الحياة ‘Living people’ and ‘أشخاص_من_أصفهان‘ people from Isfahan’. The concept Bill Gates was assigned to the categories such as Living people, American billionaires, American technology writers, American inventors, and American investors. Due to the words people, billionaires, writers, inventors, and investors are in the trigger words of the named entity type person. This concept is classified as the person. For the location, the trigger words include terms that are related to the places such as, cities, countries, village, rivers, and capitals. The NE type organization has been identified by several terms such as companies, corporation, association, union, and institution. For the events, the trigger words cover the terms referred to the events such as wars, matches, championships, revaluations, elections, festivals, parties, and invasions. Therefore, the following concepts World War II, the Japanese invasion of Manchuria, Mukden Incident, Berlin Blockade, Aden Emergency, and Yemen's revolution.

III. Result And Discussion

In sum, Table 3 reveals that most of the reported works using rule-based approach focus on MSA. This is because MSA is the current type of Arabic that is widely used nowadays. It is obvious that classical Arabic has received almost negligible attention by researchers even though it’s an important research direction due to its huge involvement with Islamic religious texts. Moreover, due to the lack of resources in the Arabic language, most of the researchers use the same corpus as can be seen from the summarizing table below. This, in fact, has restricted the research focus on few domains (e.g. political, economic) while neglecting others (e.g. medical, religious). On the other hand, the techniques used in the reported rule-based systems focus more on list look-up approach such as using gazetteers and dictionaries and less depend on rules that can be extracted from the Arabic language. The reason behind this is the complexity of Arabic grammars itself and, hence, more efforts are needed to eliminate this barrier by more rigorous studies that can come up with new rules to develop the rule-based systems’ performance. The evaluation results of the knowledge-based approaches showed that the rule-based techniques are efficient and they can improve the natural langue methods such as measuring semantic compositionality [32] and mapping lexical sources [13].

TABLE III
Summary of Literature Review for Rule-Base System

Author	Linguistic resource	Entity type	Domain	F- measure
Maloney & Niv, (1998)	TAGARAB	Person, Organization, Location, Number and Time.	Political, /MSA	85%
Mesfar, (2007)	NooJ linguistic environment	Person, Location, Organization, Currency, and Temporal expressions.	Political, /MSA	87%
K. Shaalan & Raza, (2007)	ACE and Treebank Arabic datasets	Person	Political, economic/MSA	87.5%
K. Shaalan & Raza, (2008); K. Shaalan & Raza, (2009)	many resources to build their own corpora, Treebank,	Person, Location, Organization, Date, Time, ISBN, Measurement, Filenames, Phone Numbers and Price	Political, economic/MSA	85.58%
Traboulsi, (2009)	ArabiCorpus	Time, Date and Address expressions	Political, economic / MSA	No result
Al-Shalabi et al. (2009)	many resources from newspaper	focuses on Person NEs	Political/ MSA	86.1%
Elsebai et al., (2009)	ANERcorp	Person	Political, economic/MSA	89%
Zaghouani, (2012)	ANERCorp	Person, Location, Organization	Political, economic/MSA	67.13%
M. Asharef, N. Omar, M. Albared (2012)	small crime corpus	Person, Location, Organization, data, time	Crime/ MSA	89.46%
Aboaoga and Ab Aziz, (2013)	In-house corpus collected from archives of Arabic news	Person	Political, economic, Sport/ MSA	91.71%
Hala Elsayed, Tarek (2015)	EASC corpus	Person name, Title, Countries, cities, Nationality, Date and Time	General MSA	84%

Saif, et al. [13]	Wikipedia	Person, Location, Organization, Events and Media (movies name, songs, video clips, series)	General MSA	NEs were used for enhancing mapping technique

IV. Conclusion

The progress of Arabic language processing research is still in its amateur stage compared with other different languages such as English. The reasons beyond that are the challenges inherited to the Arabic language itself together with the lack of annotated corpora and resources. It could be reported that the works on rule-based NER systems have good progress so far. However, more research efforts are still needed because most of the proposed rule-based systems are on MSA, whereas CA has received almost negligible attention. Besides, CA is almost related to religion domain and ancient Arabic literature including poetry, drama, and novels which constitute a significant research direction for Arabic NER. Hence, there is a necessity for developing new handcrafted rules that can utilize the grammars involved with the Arabic language to enhance the performance. Furthermore, the works on rule-based NER for MSA texts is still limited to few NEs types and even few domains. Therefore, there is more need to put research efforts to develop new rule-based NER systems that can introduce NER in new domains such as crime, sports, religion, etc.

References

[1] G. R. Krupka and K. Hausman, "IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7," in Proceedings of MUC, 1998.

[2] R. E. Salah and L. Q. binti Zakaria, "A Comparative Review of Machine Learning for Arabic Named Entity Recognition," International Journal on Advanced Science, Engineering and Information Technology, vol. 7, 2017.

[3] K. Shaalan, "A survey of arabic named entity recognition and classification," Computational Linguistics, vol. 40, pp. 469-510, 2014.

[4] B. Alshaikhdeeb and K. Ahmad, "Biomedical Named Entity Recognition: A Review," International Journal on Advanced Science, Engineering and Information Technology, vol. 6, 2016.

[5] UN official languages. Available: http://Un.org . Retrieved 2013-04-20

[6] UNESCO. (2014). World Arabic Language Day. Available: http://bit.ly/2lwRFYt. 18 December 2012. Retrieved 12 February 2014

[7] Executive Summary. Available: http://www.pewforum.org/2011/01/27/the-future-of-the-global-muslim-population/ . The Future of the Global Muslim Population. Pew Research Center. Retrieved 22 December 2011

[8] M. Albared, N. Omar, M. J. A. Aziz, and M. Z. Ahmad Nazri, "Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model," in Rough Set and Knowledge Technology: 5th International Conference, RSKT 2010, Beijing, China, October 15-17, 2010. Proceedings, J. Yu, S. Greco, P. Lingras, G. Wang, and A. Skowron, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 361-370.

[9] A. M. Saif, "An automatic collocation extraction from Arabic corpus," Journal of Computer Science, vol. 7, p. 6, 2011.

[10] A. M. Saif and M. J. Ab Aziz, "An automatic noun compound extraction from Arabic corpus," in 2011 International Conference on Semantic Technology and Information Retrieval, 2011, pp. 224-230.

[11] A. Saif, M. J. Ab Aziz, and N. Omar, "Evaluating knowledge-based semantic measures on Arabic," International Journal on Communications Antenna and Propagation, vol. 4, pp. 180-194, 2014.

[12] M. A. Attia, "Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation," University of Manchester, 2008.

[13] A. Saif, M. J. Ab Aziz, and N. Omar, "Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features," Natural Language Engineering, vol. FirstView, pp. 1-39, 2015.

[14] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum entropy," in Computational Linguistics and Intelligent Text Processing, ed: Springer, 2007, pp. 143-153.

[15] S. Strassel, A. Mitchell, and S. Huang, "Multilingual resources for entity extraction," in Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15, 2003, pp. 49-56.

[16] B. Farber, D. Freitag, N. Habash, and O. Rambow, "Improving NER in Arabic Using a Morphological Tagger," in LREC, 2008.

[17] M. Maamouri, A. Bies, H. Jin, and T. Buckwalter, "The Penn Arabic Tree Bank," Computational Approaches to Arabic Script-Based Languages: Current Implementations in Arabic NLP. CSLI NLP Series, 2003.

[18] S. Mesfar, "Named entity recognition for arabic using syntactic grammars," in Natural Language Processing and Information Systems, ed: Springer, 2007, pp. 305-316.

[19] K. Shaalan and H. Raza, "Arabic named entity recognition from diverse text types," in Advances in Natural Language Processing, ed: Springer, 2008, pp. 440-451.

[20] A. Elsebai, F. Meziane, and F. Z. Belkredim, "A rule based persons names Arabic extraction system," Communications of the IBIMA, vol. 11, pp. 53-59, 2009.

[21] T. Buckwalter, "Buckwalter {Arabic} Morphological Analyzer Version 1.0," 2002.

[22] J. Maloney and M. Niv, "TAGARAB: a fast, accurate Arabic name recognizer using high-precision morphological analysis," in Proceedings of the Workshop on Computational Approaches to Semitic Languages, 1998, pp. 8-15.

[23] K. Shaalan and H. Raza, "Person name entity recognition for Arabic," in Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007, pp. 17-24.

[24] K. Shaalan and H. Raza, "NERA: Named entity recognition for Arabic," Journal of the American Society for Information Science and Technology, vol. 60, pp. 1652-1663, 2009.

[25] H. Traboulsi, "Arabic named entity extraction: A local grammar-based approach," in IMCSIT, 2009, pp. 139-143.

[26] R. Al-Shalabi, G. Kanaan, B. Al-Sarayreh, K. Khanfar, A. Al-Ghonmein, H. Talhouni, et al., "Proper noun extracting algorithm for Arabic language," in International conference on IT, Thailand, 2009.

[27] W. Zaghouani, "RENAR: A rule-based Arabic named entity recognition system," ACM Transactions on Asian Language Information Processing (TALIP), vol. 11, p. 2, 2012.

[28] M. Aboaoga and M. J. Ab Aziz, "Arabic person names recognition by using a rule based approach," Journal of Computer Science, vol. 9, p. 922, 2013.

[29] H. Elsayed and T. Elghazaly, "A Rule-Based Entities Recognition System for Modern Standard Arabic," International Journal of Computer Science Issues (IJCSI), vol. 12, p. 119, 2015.

[30] M. OUDAH and K. SHAALAN, "NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic," Natural Language Engineering, pp. 1-32, 2016.

[31] A. Saif, M. J. Ab Aziz, and N. Omar. (2013, Measuring the Compositionality of Arabic Multiword Expressions. Soft Computing Applications and Intelligent Systems, 245-256.

[32] A. Saif, M. J. Ab Aziz, and N. Omar, "Measuring the compositionality of Arabic multiword expressions," in Soft Computing Applications and Intelligent Systems, ed: Springer, 2013, pp. 245-256.