NAMED ENTITY RECOGNITION IN CLASSICAL ARABIC DOMAIN USING A HYBRID APPROACH

RAMZI ESMAIL MOHAMMED SALAH

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITY KEBANGSAAN MALAYSIA
BANGI

2018

PENGECAMAN ENTITI NAMA DOMAIN ARAB KLASIK MENGGUNAKAN PENDEKATAN HIBRID

RAMZI ESMAIL MOHAMMED SALAH

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

30th of April 2018

RAMZI ESMAIL SALAH
P70146

ACKNOWLEGMENT

First and foremost, praise be to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this PhD research. Acknowledgment

I would like to express my special appreciation and thanks to my supervisor Dr. Lailatul Qadri binti Zakaria. She has been a tremendous mentor for me. I would like to thank her for encouraging my research and for allowing me to grow as a research scientist. Her advice on my research as well as on my career has been invaluable.

I would also like to thank all my friends who have supported me in writing and inspired me to strive towards my goal, and thanks too to all the postgraduate students of the UKM power research group for their help and friendship, and for creating a pleasant working environment throughout my years at UKM.

Also, a special thanks to my family. Words cannot express how grateful I am to my parents, brothers and sisters for all of the sacrifices that they made on my behalf.

In particular, I would like express appreciation to my wife, my son and my daughters who spent sleepless nights with me and were always there to give me support in the challenging periods of my study.

ABSTRACT

Named Entity Recognition (NER) is a task to identify and classify Named Entities (NEs) in unstructured language texts. Due to the different characteristics and issues associated with Arabic such as, the absence of capitalization, complex morphology and lack of structured resources, NER in Arabic has received special attentions by researchers in computational linguistics and artificial intelligence fields. Arabic NER approaches are classified into two main approaches: rule-based and machine learning approaches. The rule-based approaches use the grammatical, morphological, syntactic, and semantic information to identify the NEs. However, they are unable to deal with the complex structures of Arabic. On the other hand, the machine learning approaches depend mainly on the quality and quantity of the tagged corpus which requires manual efforts. However, the main issue associated with Arabic NER using machine learning approach is the lack of the tagged corpus especially in a specific domain such as Classical Arabic (CA) domain. Furthermore, the word-level representation of the tagged corpus usually leads to high dimensional features in the machine learning approaches. Therefore, this thesis proposed a hybrid approach for classical Arabic NER which depends on rule-based and machine learning approaches. Firstly, a new tagged corpus called Classical Arabic Named Entity Recognition Corpus (CANERCorpus) is introduced as a knowledge source. CANERCorpus contains 20 named entities that are related to CA domain such as Allah, Prophet, Paradise, Hell, and Religion. It is freely available and verified annotated by CA domain experts, containing more than 7,000 Hadiths from Sahih Bukhari. Secondly, the rule-based method has been proposed to improve and integrate the gazetteer, patterns, grammars, trigger words and blacklist for identifying Arabic named entities in Arabic contexts. Finally, the hybrid method is proposed to improve the performance of the supervised and rule-based methods. For the proposed method, a new feature called context-based feature is proposed as a novel feature to represent each type of named entity. To overcome the high dimensionality issue in the context-based representation, the genetic-based algorithm is adapted as a feature selection technique to choose only informative features from the representation of documents in the given corpus. For evaluation, the proposed methods are evaluated in CA domain using the CANER Corpus. The overall performance of the rule-based method in terms of Precision, Recall, and F-measure is 90.2%, 89.3%, and 89.5%, respectively. On the other hand, the Precision, Recall, and F-measure using the machine learning (Naïve Bayes) method is 86%, 75% and 80% respectively. Experimental results showed that the proposed hybrid method overcomes the limitations of the rule-based and machine learning methods individually. The overall Precision, Recall, and F-measure using the hybrid method is 0.933%, 0.927% and 0.930% respectively. In short, this study offers comprehensive information and evaluation regarding NER in CA domain. It opens the door for scholars and researchers to benefit the proposed methods in more complex natural language processing applications in classical Arabic domain.

ABSTRAK

Pengecaman Entiti Nama (PEN) merupakan aktiviti mengenalpasti dan mengkelas entity nama dalam teks yang tidak berstruktur. Perbezaan ciri dan isu yang berkaitan dengan bahasa Arab seperti tidak menggunakan kaedah penulisan dengan huruf besar, morfologi yang komples dan kekurangan sumber yang berstruktur dalam PEN Arab menjadi perhatian penyelidik dalam bidang komputeran linguistik dan kecerdasan buatan. Pendekatan PEN Arab dibahagi pada dua iaitu berasaskan peraturan dan pembelajaran mesin. Pendekatan berasaskan peraturan menggunakan maklumat tatabahasa, morfologi, sintaktik dan semantik untuk mengenal entiti nama. Namum begitu, ia tidak dapat menangani isu struktur bahasa Arab yang kompleks. Manakala, pendekatan berasaskan pembelajaran mesin bergantung sepenuhnya pada kualiti dan kuantiti korpus yang telah ditag secara manual. Namun begitu, kekurangan korpus bertag merupakan isu utama yang dikait dengan NER Arab. Seterusnya, perwakilan aras perkataan dalam korpus bertag lazimnya membawa kepada sifat dimensi yang tinggi dalam pendekatan pembelajaran mesin. Oleh itu, tesis ini mencadangkan pendekatan berasaskan hibrid bagi PEN Arab klasik yang menggunakan pendekatan berasaskan peraturan dan pembelajaran mesin. Pertama, korpus baharu bernama Korpus Pengecaman Entiti Nama Arab Klasik (CANER Corpus) diperkenal sebagai sumber pengetahuan. CANER Corpus mengandungi 20 jenis entiti nama yang berkaitan dengan domain Arab klasik seperti Allah, Nabi, Syurga, Neraka dan Agama. Ia boleh didapati secara percuma dan anotasinya telah ditentusah oleh pakar bidang Arab klasik (CA) dan mengandungi 7000 hadis daripada Sahih Bukhari. Kedua, kaedah berasaskan peraturan dicadang untuk menambahbaik dan mengintegrasi teknik gazetir, pola, tatabahahasa, pemicu perkataan dan senarai hitam untuk mengenalpasti entiti nama Arab. Ketiga, ciri baharu bernama ciri berasaskan konteks telah dicadang sebagai ciri novel dalam kaedah pembelajaran mesin. Untuk menangani isu demensi yang tinggi, algoritma berasaskan genetik telah diguna untuk memilih ciri yang berinformatif dari perwakilan dokumen dalam korpus tersebut. Akhirnya, metod hibrid diimplementasi untuk menambahbaik prestasi metod berasaskan peraturan dan pembelajaran mesin. Kaedah tersebut dinilai dalam domain Islam dengan menggunakan CANERCorpus. Prestasi keseluruhan kaedah berasaskan peraturan diuji dengan analisis kejituan, dapatan semula dan ukuran-F dan hasil masing-masing adalah 90.2%, 89.3%, dan 89.5%. Seterusnya, kejituan, dapatan semula dan ukuran-F menggunakan pendekatan pembelajaran mesin (Naïve Bayes) adalah masing-masing 86%, 75% and 80% Hasil eksperimen menunjukkan pendekatan berasaskan hibrid yang dicadang telah mengatasi kekurangan pendekatan berasaskan peraturan dan pembelajaran mesin. Tesis ini telah menyediakan maklumat dan pengujian yang menyeluruh berkaitan dengan PEN dalam domain Arab klasik. Hasil kajian ini diharap akan membuka ruang dan memberi faedah kepada penyelidik lain mengkaji dokumen Arab klasik dan membangun aplikasi pemprosesan bahasa tabii.

TABLE OF CONTENTS

LIST OT TABLES

LIST OF FIGURES

Figure No. **page**

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
IE	Information Extraction
IR	Information Retrieval
NE	Named Entity
NEs	Named Entities
NER	Named Entity Recognition
ANER	Arabic Named Entity Recognition
CANER	Classical Arabic Named Entity Recognition
CANERCorpus	Classical Arabic Named Entity Recognition Corpus
MUC	Message Understanding Conference
MUC6	The Sixth Message Understanding Conference
CA	Classical Arabic
MSA	Modern Standard Arabic
ML	Machine Learning
SL	Supervised Learning
SSL	Semi Supervised Learning
SVM	Support Vector Machines
CRF	Conditional Random Field
NB	Naïve Base
HMM	Hidden Markov Model
ME	Maximum Entropy
ANN	Artificial Neural Network

CHAPTER I

INTRODUCTION

1.1 BACKGROUND OF THESIS

This chapter introduces the thesis, which investigates Arabic Named Entity Recognition (ANER). The ambitious goal of Natural Language Processing (NLP) is to have machines interact with the languages in the way that humans do. In other words, the computers should be able to understand the context constructed according to the grammar of some NLP and should be capable of generating in reply meaningful sentences in this language. The NLP is a field of computer science which it can be considered as a sub-area of artificial intelligence and inherits from it many theories and techniques.

Information Extraction (IE) is one application of NLP which automatically extracts structured information such as entities, relationship between entities and also attributes describing those entities from unstructured documents (Cowie & Lehnert 1996). IE systems are effective for tackling the problem of information overload as they enable us to get the most important part of information that exists in huge documents quickly and easily.

Named Entity Recognition (NER) is the subtask of IE that detects and classifies proper names, which are called Named Entities (NEs), within unstructured texts into predefined types such as Persons, Locations or Organizations, etc. (Nadeau & Sekine 2007; Shaalan 2014). Other than IE, NLP applications such as those mentioned above make use of NER. For example, in Information Retw3rieval (IR), NER can be used first to recognize NE in the query, and to extract relevant documents containing this NE. In Machine translation recognizing NEs is important in order to disambiguate some words

from NEs. Finally in Question and answering NER plays the same role as in IR (Oudah & Shaalan 2012; Shaalan 2014).

In 1990, at the Message Understanding Conferences (MUC) introduces the task of NER and was given attention by the community of researchers. These conferences were funded by the Defense Advance Research Project Agency (DARPA) for the purpose of developing a better information extraction system. At the sixth MUC, three main NER subtasks were defined, namely: ENAMEX (i.e. Person, Location and Organization), TIMEX (i.e. temporal expressions), and NUMEX (i.e. numerical expressions) (Oudah & Shaalan 2012; Shaalan 2014).

In relation to the above, a named entity refers to a term/word used to identify an object from object groups possessing similar traits. Named entity is a term first proposed at the Message Understanding Conference (MUC-6) back in 1995 (Grishman & Sundheim 1996). In the named entity expression, the word named confines the scope of entities that have a single or several rigid designators standing for referents. Rigid designators encompass proper names but are dependent on the domain under examination that often times relate the reference word for object in the domain as named entities. This can be exemplified by the entities of interest in the molecular biology and bio-informatics field that are genes and gene products. In prior studies, the dominant classification of NE e.g., in (Nadeau & Sekine 2007; Shaalan & Oudah 2014; Saif et al. 2015) encompasses three classes namely names of person, location and organization. Nevertheless, there are many specific types for NE that arises in specific domains such as diseases and medicine in the case of biomedical domains as illustrated by (She et al. 2015)

1.2 ARABIC NAMED ENTITY RECOGNITION

As with the English language, the ability to effectively identify NE for Arabic is important owing to its importance as a factor in majority of NLP applications. For the English language, NER has been exhaustively examined and but further work is still needed for Arabic.

Arabic is described as a Semitic language that produced morphological and orthographic challenges such as the way proper names are considered as common language words, the absence of capitalizations and conjunctions, prepositions, possessive pronouns, and determiners connected to words in the form of pre-fixes or suffixes (Abdul-Hamid & Darwish 2010). The following statements summarize the key challenges;

- To identify new grammar rule that assist in increasing performance.
- To identify a set of features that functions well for Arabic NER.
- To develop a new way to pre-process Arabic text (how to deal with morphology, etc.).
- To determine an effective approach to identify and categorize NE.

Arabic is considered as the official language of citizens of the Arab world, with over 300 million people having Arabic as their mother-tongue (Shaalan 2010). Added to this, Arabic is a Semitic language and is the language of the Holy Quran and as such, every Muslim, worldwide makes use of Arabic in their daily prayers. Arabic, like other Semitic languages, is a rich natural language as represented by its morphology and inflection (Aljasser & Vitevitch 2017).

In the past few years, NLP for Arabic has garnered increasing attention in light of its challenges (Oudah & Shaalan 2012), particularly in extracting information because of the complexity of the language stemming from its rich morphology. Nevertheless, the Arabic NER is still in its infancy and opportunities abound for improving its performance. Several NER systems in Arabic were proposed, as mentioned through the primary use of two types of approaches namely the rule-based approach, and ML-based approach. To maintain rule-base systems consumes considerable labor and significant amount of time, particularly with the lack of linguists' availability. Contrastingly, the ML-based NER systems make use of ML techniques that need robust tagged datasets for the purpose of training and testing. In other words, ML-

based NER systems are employable and updatable with the least time and effort so long as the datasets are robust. To this end, the lack of linguistic resources leads to ample hindrance in Arabic NLP, particularly Arabic NER.

ANER systems have been reported to face some issues related to Arabic language, like capitalization (Shaalan 2014), complexity in morphology (Atwan et al. 2016), and minimal resources (Saif et al. 2015). In other studies, Arabic NEs have been introduced through three main methods namely, rule-based method by (Benajiba et al. 2009; Zaghouani 2012), machine learning by (Benajiba et al. 2008; Mohammed & Omar 2012) and hybrid approaches by (Oudah & Shaalan 2012; Shaalan & Oudah 2014). The major issue faced by the supervised Arabic NE recognition includes knowledge elicitation bottleneck and lack of resources to tackle underdeveloped languages that calls for extensive efforts from linguist circles.

A review of literature shows that studies dedicated to Arabic NER are few and far between and hence, more investigation should be conducted on the topic to resolve its challenges, particularly in Hadith and Quran.

1.3 MOTIVATIONS

Generally, several NLP applications such as machine translation, information retrieval and question answering that rely on NER as a pre-requisite stage. Literature reports three types of approaches used for the development of NER systems and they are handcrafted rule-based approach, machine learning (ML) based approach and lastly, hybrid approach. First, the rule-based approach is dependent on handcrafted grammatical rules, ML-based approaches leverages different ML algorithms using sets of features obtained from annotated datasets (annotated with named entities) for NER system development. Lastly, the hybrid approach is a combination of the two former approaches to enhance the overall NER system performance. In fact, NER has been employed on various natural languages like English, French, German, Chinese and Arabic.

In Arabic language processing, MSA is invaluable in dealing with day-to-day

news and sports and social media (Shaalan 2014; Aljasser & Vitevitch 2017), and CA is interlinked to the Islamic domain that is important to Arabic and the whole Islamic world. Stated clearly, Arabic language is extensively spoken and it is considered to be an influential language. It is the official language of the Gulf countries and the language of the Holy Quran and thus, Muslims all over the world, constituting $1/6^{th}$ of the world population, has religious affection towards it. The Gulf region is rich in light of natural resources like oil, natural gas and other minerals, and this makes the language of Gulf, which is Arabic, important in geographic and political sense. In this regard extraction of information is important in Arabic language.

In the Arabic text, NLP is relevant as Arabic is a language that is spoken by over 300 million people around the globe and because there are considerable number of Arabic sites online. Owing to the different features of the Arabic language, its complexity and rich morphology and the variation in orthographic aspects as well as the non-capitalization of Arabic texts, NLP is in this context is difficult. To the best of the researcher's knowledge, many NER systems integrate gazetteers with rules, considering elements in the overall context. In the first approach to recognize named entities from the Arabic text, the author decided to employ a gazetteer lookup – where a gazetteer refers to a list of named entities that are known.

In this background, NER is used for different tasks, aside from extracting information. It is also useful for machine translation, searching results, clustering and question/answers, among others. Identifying NEs can be utilized as a pre-requisite process for several NLP Systems. NER for Arabic language is still in its infancy and thus, little work has been done for Arabic NLP and NER. This lack in studies can be attributed to the lack of available tools and language resources for Arabic language that are linked to named entity recognition task. It may also be attributed to the challenge in dealing with Arabic language in terms of linguistic grammar-based processing and some of these challenges were highlighted by in their study.

Literature abounds with studies focused on MSA and thus, focusing on CANER motivates this study as only a few have been conducted on the latter, particularly relating to Islamic texts.

1.4 PROBLEM STATEMENT

Due to the unavailability capital letters and morphological complexity of Arabic, the Arabic NER is a special task that has received wide attention by introducing different techniques to handle these issues. In other words, Arabic NER systems are facing some challenges that are associated with Arabic language such as, capitalization issue (Shaalan 2014; Zirikly & Diab 2014; Alanazi 2017), complex morphology (Oudah & Shaalan 2012; Shaalan 2014; Atwan et al. 2016) and lack of resources (Oudah 2012; Aboaoga & Ab Aziz 2013; Shaalan 2014; Zaghouani 2014). Several studies have been introduced for Arabic named entities using three main approaches, rule-based (Mesfar 2007; Elsebai et al. 2009; Halpern 2009; Traboulsi 2009; Shaalan 2010; Zaghouani 2012; Aboaoga & Ab Aziz 2013; Elsayed & Elghazaly 2015), machine learning (Benajiba & Rosso 2007; Benajiba et al. 2007; Benajiba & Rosso 2008; Abdul-Hamid & Darwish 2010; Zirikly & Diab 2014; Al-Shoukry & Omar 2015; Althobaiti et al. 2015; Dahan et al. 2015; Zirikly & Diab 2015) and hybrid approaches (AbdelRahman et al. 2010; Abdallah et al. 2012; Oudah & Shaalan 2012; Meselhi et al. 2014; Meselhi et al. 2014; Shaalan & Oudah 2014; Alanazi 2017). Previously, these approaches have been extensively evaluated in the modern standard Arabic. However, to the best knowledge of the author, there is lack in the investigation of NER in classical Arabic which is closely related to Islamic domain (Shaalan 2014). Therefore, using classical Arabic for conducting NER experiments can provide valuable insights about the NER task as well as applying the exiting approaches to a new domain (Islamic) of classical Arabic.

By adapting conventional NER approaches to classical Arabic, the recognition of Arabic NEs in the text faces many serious problems that spread over all the levels of processing from the first stage until the final stage of extraction. These problems are determined in each stage of the recognition NEs. Rule-based approaches for NER rely on linguistic knowledge to detect and classify named entities. These approaches use the morphological, syntactic, and semantic information to introduce different rules that recognize the NEs. However, due to lack of classification rules/guideless knowledge (Abdallah et al. 2012; Oudah 2012; Aboaoga & Ab Aziz 2013; Shaalan 2014; Zaghouani 2014), and inaccurate rules, they are unable to deal with the complex

structures of NEs (Karaa & Slimani 2017). In addition, the rules that have been proposed for modern standard Arabic are insufficient to recognize NEs in the classical Arabic especially Islamic contexts that have different forms from the modern standard Arabic such as lexical meaning, special symbols, absent of dots (Shaalan 2014; FAIZAL et al. 2015; Aljasser & Vitevitch 2017) . Moreover, the linguistic rule approaches cannot generate some kinds of NEs in classical Arabic such as the Allah (names of god in Islam), the prophet (names of Allah's messengers), and hell (places of punishment in an afterlife in Islam beliefs).

On the other hand, supervised-based approaches utilize different learning algorithms to generate statistical models for NE prediction (Benajiba et al. 2007; Benajiba et al. 2008; Oudah & Shaalan 2012; Meselhi et al. 2014; Zirikly & Diab 2015; Alanazi 2017; Aljasser & Vitevitch 2017; Karaa & Slimani 2017). One of the main problems facing the supervised approaches is the knowledge acquisition bottleneck in the classical Arabic (Shaalan 2014). The knowledge for the supervised approach is the tagged corpus which typically requires domain experts who can recognize each named entity in the given corpus (Bontcheva et al. 2017). In addition, these approaches are unable to deal with non-informative features. Moreover, the supervised approaches suffer from the high dimensionality issue resulting from representation of features especially when the tagged corpus is very large. In the hybrid approaches, the serious challenge in recognizing NEs is to construct logically sound models for describing the different types of NEs, because the NEs have the high variety of linguistic features and include a wide range of phenomena that are related to extract and select the features.

Due to the several limitations mentioned, recent approaches suffer from all the above issues in terms of extract named entity from classical Arabic.

1.5 RESEARCH QUESTIONS

This research is set up to answer the problem statement related questions, such questions are formulated based on the research problems that should be investigated and solved during this research. The following questions are the core of our research and it should be answered throughout this research:

RQ1: How can new types of classical Arabic NER corpus overcome the knowledge acquisition bottleneck problem in the classical Arabic?

RQ2: How can the rules-based including gazetteers, trigger words, black list, patterns and grammatical forms are integrated together towards improve the performance of the NERs in classical Arabic?

RQ3: How can the advantages of both rule-based approach and supervised approach are incorporated and how can feature extraction and selection be employed to handle the high dimensionality issue?

Thus, the hypothesis of the research can be stated in the form of claim as:

"Hhybrid approach can overcome the limitations of the rule-based and supervised approaches ".

1.6 RESEARCH OBJECTIVES

The objectives of a research study summarize what is to be achieved by the study. These objectives should be closely related to the research problem. Thus, based upon the problem previously discussed and the above three research questions, the objectives of the study are as follows:

RO1: To identify new types of NE for CA by building a classical Arabic NER corpus. This objective is dedicated to identify a new NEs and build classical Arabic entity recognition corpus to be used as a dataset for the current study.

RO2: To develop a rule-based method to identify new types of NE for CA and enhance the performance. This objective is for new NEs and enhance other types by integrating gazetteers, trigger words, black list, patterns and grammatical forms to identify the Arabic named entities to improve the performance of NERs in classical Arabic.

RO3: To propose a new hybrid approach (The integration between the machine learning and the rule-based analysis) to handle the high dimensionality issue and overcome the limitations of the rule-based and machine learning.

In the relation to the three-research objective, which attempts to develop the hybrid approach for ANER, there is an evaluation for each approach of the study. The present study attempts to create and propose a hybrid NER system for Arabic that can extract 20 different types of named entities (person, location, organization, measurement, money, crime, natural object, book, date, time, month, day, number, Allah (God), Prophet, paradise, hell, religion, sect, clan, among others). The proposed system comprises a combination of the rule-based component and the ML-based component, with the former being a reproduction of a prior rule-based NER system by Shaalan and Raza (2008), having certain modifications and additions to improve the system's performance and accuracy. Meanwhile, the ML-based component uses ML techniques that were successfully utilized in similar NER of other languages. It is used to produce an Arabic NER model upon an annotated dataset generated by the rule-based component. Moreover, the annotated dataset is introduced to the ML-based component using a set of features that are carefully and logically chosen for the performance optimization of the component. As for the linguistic resources, two types are gathered and obtained as required namely, gazetteers and corpora (datasets). The data undergoes verification and preparation stages prior to the NER application. Different experiments are carried out for evaluating the proposed system based on different dimensions and its output is exploited to recommend new grammatical rules that may enhance the rule-based component's performance.

1.7 RESEARCH SCOPE

This research is focused on a CANER corpus, and it employs a methodology to resolve the issue of a specific domain. The focus of this research is placed on a specific issue of NER in the domain of Arabic language and Hadith. It does not attempt to solve the entire issues of NER relating to Arabic language but is confined to a definite scope. This study is limited to the following scope;

- This research is focused on a specific language, namely Arabic, specifically Classical Arabic (CA), which the language spoken by the Arabic in medieval history.

- This research is also limited with regards to the examined domain, which is the Islamic text (Hadith) domain. Hence, the rules and technical related to the development of this specific domain as applied to the system may differ from those using it in the development of another domain.

- With regards to the types of NE, this study is limited to the most famous 20 different NE types namely, person, location, organization, measurement, money, crime, natural object, book, date, time, month, day, number, Allah (God), Prophet, paradise, hell, religion, sect and clan. These 20 types are more related to CA and have more occurrences and frequencies.

Therefore, this study has chosen to focus on developing a named entity recognition in classical Arabic domain using a hybrid approach.

1.8 SIGNIFICANCE OF THE RESEARCH

The significance of this study can be discussed in relation to theory, religion, and practice. This study is applied on the basis of the different reasons; first, different Hadith language recognition has been invaluable to individuals, specifically Muslims. In the past several decades, people all over the globe have been provided an opportunity to learn and understand the teachings of the Hadith through its translation into different languages.

Second, with the NER Arabic version of the Hadith, majority of native English speakers can interpret, study and understand the teachings of the Hadith in an easy way. This study conducts a significant examination to propose a hybrid approach that will meet the demand for an accurate NER of the classical Arabic language. By combining conventional semantic relatedness measurements (rule-based, ML approach), the task is expected to bring about the greatest performance level.

Moreover, the used approaches are expected to improve through the generation of considerable amount of semantic knowledge based on the new corpus sources, for the purpose of solving issues faced by translators. The issues are related to the use of multiple word meanings in the Arabic language. The past acknowledged aspects transform this study into an important reference that could influence further interest on in-depth examination in the field of the calculation of the accurate meaning of words and sentences. This could enhance the evaluation of the extant methods of Arabic recognition.

1.9 RESEARCH METHODOLOGY

The determination of the right methodology is one of the most significant stages of carrying out a research project. The right research method facilitates the establishment of objectives and their successful achievement to contribute to the study field. This study employs a methodology culled from past studies to minimize the research gap. The identified gaps call for solutions to be examined in a cautious manner. This is followed by the development and implementation of the solution prior to its evaluation and comparison to other related works. These are deemed to be the major steps in the research methodology.

Accordingly, this study adopts a generic system development research methodology known as the design-based research (DBR) (Barab & Squire 2004; Wang & Hannafin 2005; Herrington et al. 2007), and it is extended to achieve the research nature and needs. The DBR methodology is employed to solve the research problem through a proposed hybrid approach consisting of the rule-based approach and the machine learning approach, utilizing the CARECorpus.

In particular, this study adopts a research methodology comprising of four major steps as established by Reeves (2006) and that are presented in Chapter Three. A summarized information of these phases is explained as follows:

- **Problem Identification:** In this phase, the literature on NER and methods that were used to develop the hybrid method is reviewed. The role based and

Machine Learning approaches and their features are also discussed.

- **Data Collection:** In this phase, two sub-phases are followed; the first was dedicated to create a new corpus known as CANERCorpus. The second, investigate and analyses data in order to use features of this data in rule-based and ML methods.

- **Design and Development:** In this phase, the proposed approach is developed, namely: the hybrid method. This method is developed based on two approaches named: Rule-based and Machine learning.

- **Evaluation:** In this phase, the evaluation and measurements for each approach is presented.

1.10 RESEARCH SUMMARY

In the next Chapter, the available approaches that address NER problems in general and particularly for Arabic will be discuss and review. The reviewing process is considered as an elementary and essential phase in this research where the research framework is identified based on the observations made on the previous related works that have been conducted for ANER. As shown in Figure 1.1, three directions of research are highlighted based on the analysis of important features that include challenges, problems, and research objectives with their intended outcomes. An iterative approach is utilized around these directions and a prototype is designed, developed and evaluated for each phase as a proof of mentioned concepts in this research.

The first direction of the current thesis is to handle the lack of available knowledge base sources of Arabic. The past decade has witnessed construction of the background information resources to overcome several challenges in text mining tasks. For non-English languages with poor knowledge sources such as Arabic, these challenges have become more salient especially for handling the NLP applications that require human annotation. The second direction deals with overcoming the challenges of using Arabic resources and rules for identifying Arabic named entities. The rule-

based methods are proposed in this direction by incorporating gazetteers, white lists, and grammar rules. The third objective aims to overcome the problems of high dimensionality in the supervised methods. In this direction, the new type of feature called context-based features is introduced to represent Arabic texts. The last objective aims to overcome the limitations of the rule-based and machine learning methods. The research summary is shown in Figure 1.1.

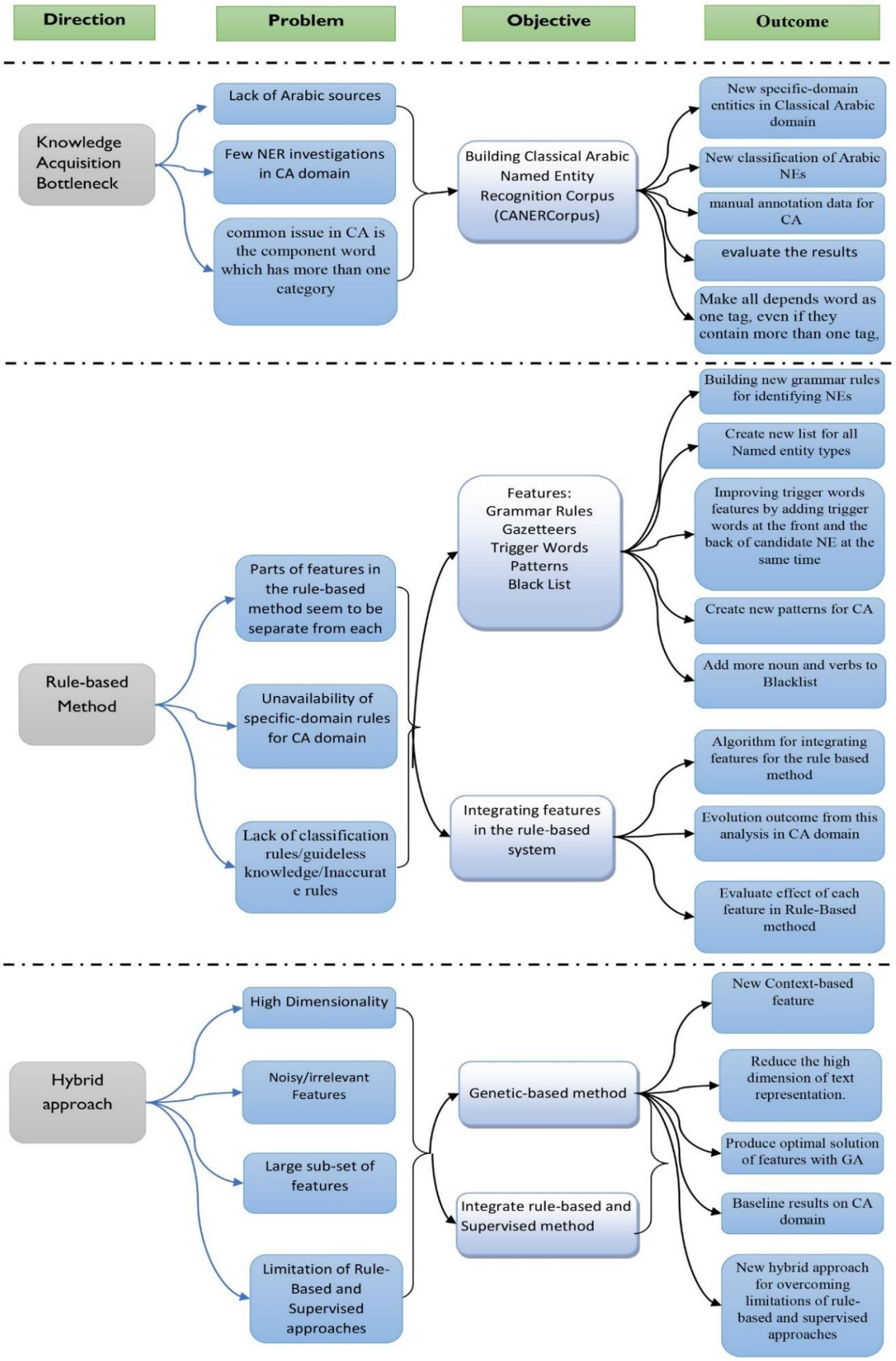

Figure 1.1 The research summary

Thus, the operators of genetic algorithm (GA) are modified to overcome limitation of high dimensionality representation of the dataset. A hybridization of feature selection measures with GA has proposed to overcome the randomization and time-consuming problems of GA and to add further improvement of feature selection process in the supervised method of ANER.

1.11 ORGANIZATION OF THE THESIS

The present research is divided into eight chapters organized as follows;

Chapter one provides the introduction of the study. This is followed by chapter two that contains the review of related studies regarding NER and the methods used and the discussion within assists in the study formulation. Chapter three contains the research methodology in detail, beginning from problem identification to meeting the study objectives.

Chapter four provides detailed descriptions of the Classical Arabic approach called entity recognition corpus (CANERCorpus), while chapter five contains an overview of the rule-based approach and discussion of outcome. The chapter also conducts an evaluation of the performance of the proposed approach. This is followed by chapter six, wherein the details of ML and its implementation are presented along with the results.

Chapter seven then presents detailed explanation of the hybrid approach for Classical Arabic called entity recognition along with the implementation outcomes. Lastly, chapter eight provides the conclusion of the thesis and it contains general and specific contributions in the three areas of study. Conclusions and recommendations are also provided in this chapter.

CHAPTER II

BACKGROUND AND LITERATURE REVIEW

2.1 INTRODUCTION

This chapter introduces the relevant concepts significant in the understanding of NER. It begins by introducing the NE and the types of Arabic language, followed by a discussion of the Information Extraction (IE). The chapter then provides a discussion of the IE components, with NER as one of them. NER is discussed in detail in terms of its architecture, used approaches, feature sets and evaluation metrics. The chapter continues to discuss the Arabic language and the challenges related to it and lastly, related studies concerning NE approaches namely rule-based, ML and the hybrid are presented.

2.2 LITERATURE MIND MAP

Literature map can help in simplifying the road maps of the whole literature and give a brief image about the research. Figure 2.1 shows the literature map of the proposed study.

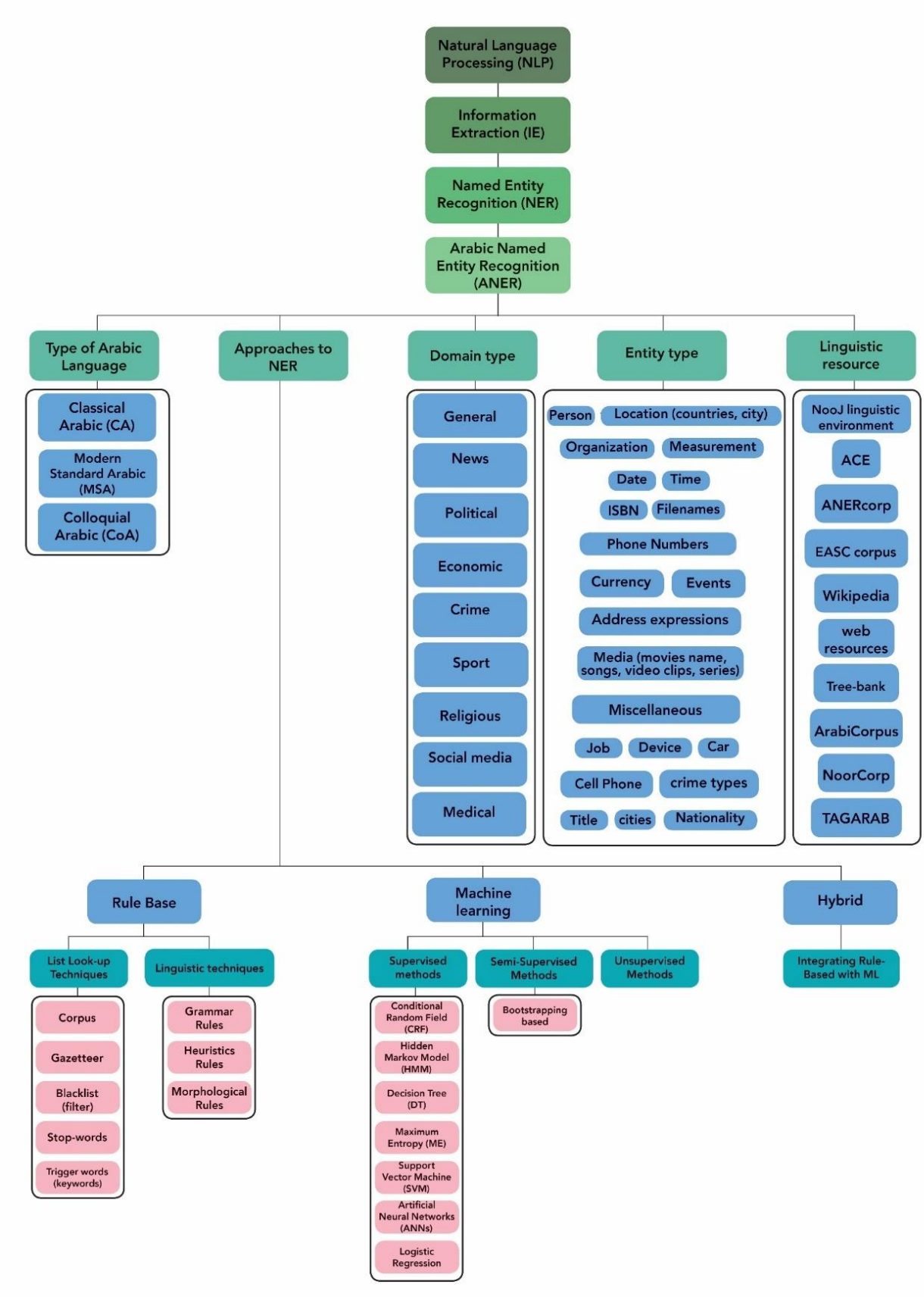

Figure 2.1 Literature mind map

2.3 PRELIMINARIES

Some preliminaries on information extraction, named entity, are presented in the following subsections.

2.3.1 Information Extraction

The increasing information growth that is accessible online has been facilitated by the Internet, which has become the global knowledge repository. Relating to this is the major issue of information overload that results from the easy manipulation, process and storage of data online. This can be solved through information retrieval that is referred to as a sub-file of Information Science. Gaizauskas and Wilks (1998) related that Information Retrieval (IR) responds to the query of the user by enlisting documents known to be relevant to the query based on their importance. It enables the user to sift through the documents to determine which are important and which are not. It also assists the users to highlight keywords in the query in the documents that need retrieval. Although the contemporary IR system offers several features enabling users to search for relevant documents, majority of IR systems are not capable of handling natural language texts aside from identifying the relevance of documents through the calculation of relative words frequency in the corpus that corresponds to the query.

Similarly, another alternative to IR systems is IE. It is also utilized to address information overload by extracting facts existing in natural text and allowing the storage of such facts in an organized form for manipulation and analysis (Grishman 1997; Jurafsky & James 2000). IE is defined by Cowie and Lehnert (1996) as the process that begins with texts collection, transformation of texts into information that can be digested and analyzed, isolation of relevant text fragments, and extraction of relevant information from the fragments. Lastly, the targeted information is formed together to form a clear framework.

In history, MUC was sponsored by DARPA in the 1980s to popularize information extraction as the tasks involved are defined by MUC. The evaluation of the IE systems was conducted through standard evaluation metrics. The next paragraphs

describe tasks defined by the MUC known as generic tasks that enable IE systems evaluation based on task performance. The tasks also facilitate consistent and reliable IE systems evaluation. The descriptions of such tasks as established by (Grishman 1997; Jurafsky & James 2000) are provided below;

Named Entity Recognition – in majority of IE systems, the first step involves detecting and classifying named entities that comprises of proper nouns within a natural text. According to the domain within which the NER application is applied, named entities can refer to several things; for example, in majority of generic news-oriented NER systems, NE types indicate places, persons and organizations. In some applications, the other entity types may have to be identified (e.g., proteins, genes, weapons, etc.). The present study is focused on NER related tasks and thus, it provides a detailed description of the task in the later part of the thesis.

Relation Detection and Classification – following the detection and classification of the NE types, IE moves on to detect and classify the relations existing between the types of NE. The generic relations existing between various entities were categorized by Jurafsky and James (2000) as follows; 1) affiliation – relating to an individual/organization or to one another (e.g., married to, spokesman for etc.); 2) geospatial – relating locations to one another (e.g., near and on the outskirts); and 3) the part of the relation relating part of something such as unit of and parent of. Majority of the relations existing in domain-specific applications exemplify the above types of relationships (Shaalan 2014).

Temporal and Event Processing – after the different NE types and their relationships are detected and classified, the IE system then searches for and analyzes the events within which the entities interact with and the way the events relating to time are significant for the extraction of information from a specific text (Imran et al. 2015).

Template Filling – this pinpoints documents invoking specific script that comprise prototypical sequences of sub-events, participants, roles, and properties, and the fills slots in the related templates with fillers directly obtained from the text. The fillers may be composed of NEs or text segments that are taken from the text (Elhadad

et al. 2015).

2.3.2 Named Entity

Named entity refers to any object that is included in a set of other objects covering similar traits. It is an expression that refers to the notion that the scope of a specific entity is confined by particular word although the entity has many rigid designators that can represent a referent. In fact, rigid designators often have proper names (Rodrigues et al. 2014). They depend on the domain of interest with the reference word related to an object within it. For example, in the case of bio-informatics and molecular biology, the interest entities are gene products as well as genes and as such NER automatically attempts to identify and classify such names in texts and categorizes them in pre-defined classes. A total of five NE included in the Arabic sentence are shown in Table 2.1.

Moreover, a named entity can be referred to as a word or term that determines an objective in a given set of objects. The entire objects within the set share similar traits. The expression NE shows that the word named limits the entities scope including one or many rigid designators that indicates a referent.

Table 2.1 Example of a named entity

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

Table 2.2 Identifying and labelling entities by the named entity extractor.

Table 2.2 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

2.4 ARABIC LANGUAGE

Arabic language is universally ranked as the 6th major language of the world (Aljasser & Vitevitch 2017). It is the official language of 22 Middle Eastern countries (UNESCO 2014) and it has a direct religious influence on over 1.6 billion world population (Lewis et al. 2016). The world's Muslim population is expected to increase by about 35% in the next 20 years, rising from 1.6 billion in 2010 to 2.2 billion by 2030, according to new population projections by the Pew Research Center's Forum on Religion & Public Life (Grim & Karim 2011), so Arabic is widely learned as a foreign language in non-Arab Muslim countries and by other Muslims all over the world

Generally speaking, there has been little development in the NLP of Arabic Language in comparison to other universal languages, topped by English as the language of Lingua franca. It was only recently that increasing studies have been dedicated to Arabic text NLP and NER.

2.4.1 Type of Arabic Language

Arabic language has three forms, namely Classical Arabic (CA), Modern Standard Arabic (MSA), and Colloquial Arabic Dialects (Shaalan 2014; Aljasser & Vitevitch 2017). Several differences can be noted between the different uses of the language types that are significant to the development of NLP systems as each NLP application aimed at a single language type will produce mixed results if used in other types. The above-mentioned forms are explained as follows;

a. Classical Arabic (CA)

For the past 1500 years or more, Classical Arabic has been utilized as the language of Islam and majority of Arabic historical and religious texts are handwritten using CA. Added to this, the extraction of Arabic NE from CA has transformed into a crucial topic, particularly when digitized CA materials are transformed from historical manuscripts. In the current times, CA is sometimes known as the Quranic Arabic, indicating the original Arab tribes' dialect in the Arabic peninsula back in the medieval ages. CA, as

a spoken language, was extensively utilized until the 9th century, during the Abassid era. Following that era, the Arab world opened up to adopt surrounding cultures and languages including Turks and Persian that has brought about changes and transformations to CA to how it is today, the Modern Arabic.

Additionally, CA is also referred to as Ancient Arabic and is a major branch of Semitic language that was deemed to be the communication tool during the advent of Islam. CA's imaginative and sophisticated nature drive those who can speak it to pride themselves. As for the CA grammar, it is quite complicated and engaging in the text, with well-structured and highly-layered vocabulary. However, regardless of its complex structure, other languages do not come close to matching CA's beauty.

Despite the fact that CA is not completely utilized as the medium of communication, it is important and significant as it is the basis of modern Arabic that is used in the current times and to master the language, CA is essential. More importantly, CA is significant as the Arabic and Islamic culture and heritage stem from the Quran (Muslims' holy book) and other Islamic and historical resources (Hadith) that are all in CA (Salah & binti Zakaria 2017).

b. Modern Standard Arabic (MSA)

The Modern Standard Arabic (MSA) is the Arabic that is extensively used in the current times as the official language of communication among governments, sectors and individuals. MSA can be described in principle as the skimmed CA version and based on its name, it is the evolved version of CA. This Arabic form is the official language of 22 Arab states used in oral and written formal interactions, occasions and events. Additionally, the UN accounted that MSA is one of the main official languages (DENNIS et al. 2012) and the most well-known Arabic language form used in education, media and magazines. Majority of the studies that introduced the Arabic language documents analysis evaluated MSA information resources like Part Of Speech (POS) tagging (Albared et al. 2010), extraction of collocation (Saif et al. 2014), extraction of noun compound (Saif et al. 2014), and measurement of semantic similarity (Saif et al. 2014). Majority of the Arabic NLP, encapsulating NER research projects,

are focused on MSA. Aside from this, the top significant difference between the two Arabic language types (MSA and CA) lies in their vocabulary, containing NEs and the conventional written Arabic orthography.

Moreover, what makes MSA for affordable is it is devoid of short vowels and it reveals the need of the up-to-date expression. Such versatility is missing in CA that often represents the older prerequisite styles. For example, the Arabic NEs in old manuscripts and documents that refer to jobs, organizations and places are different from the corresponding NEs found in contemporary documents (Attia 2008).

c. Colloquial Arabic

This Arabic form is a dialect that is specific to regions and is mostly used in oral communication and social media. Majority of the words are taken from MSA in that it is different from one area of the Arab world to the next. For example, عبدالقادر (Abd Kader) compared to عبدالجادر (Abd Al-Gader) or عبدالآدر (Abd Al-Aader) (Shaalan 2014; Zirikly & Diab 2014).

2.4.2 Differences between Classical and Modern Arabic

The distinct nature of the Arabic language lies in its two primary types, which are CA and MSA. CA is the basis of modern Arabic. While it involves modern Arabic, it displays specific features in comparison to MSA. It is thus impractical to assume that both types are identical when developing an NER system. Some ideologists might refute the distinct difference between MSA and CA but when examining both, significant differences are notable in their lexical meanings, style and some constructions of grammar. Some of the significant differences between MSA and CA are discussed in this section.

a. Lexical meanings

There are many words in CA that are not used in MSA or that relay different meanings. For instance, the word نائبة in CA relays the meaning of catastrophe, whereas in MSA its

meaning can be the name of a female person (Naeba- a female deputy). This is also exemplified by the idiom in CA أَنتَ تَقِيٌّ وأَنا مَقِيٌّ فكيف نَتَّفِق which means (Anta Taaek Wa-Ana Maaek fakaif Natafik – you are hot-tempered and I am weepy, so how can we be in harmony), but the words تَقِيٌّ (hot-tempered) and مَقِيٌّ (weepy) are not in use in MSA in contemporary times.

b. Special symbols

Under this sub-title, some specific CA symbols and diacritics are not used in MSA. Therefore, one NER application focused on MSA will generate mixed results when focused on CA. Majority of these specific symbols are used in the Holy Quran to aid readers in identifying text clauses, beginning/end of verses (ayats) and determining the functionality of Arabic alphabet in one word. Therefore, they can deemed as auxiliary morphological and lexical symbols and their absence can result in the ambiguity of NER. For instance, the word فَصَبَرُوا, the last alphabet in the word indicates ا aleph with circle symbol on top – the symbol signifies an extra letter, in that the actual word is فَصَبَرُو, which translates to being patient.

c. Absent of dots on some old Arabic manuscript

In some ancient Arabic manuscripts, the familiar dots related with the Arabic alphabet are missing. For instance, the letter ن is written without a dot on top. The Arabic language characteristics are challenging to deal with owing to their peculiarities and distinct nature.

2.4.3 Challenges of Arabic Language

Several studies have been conducted on NER of Latin-scripted languages including English, German, Spanish, and Dutch, with only a few studies on Arabic text (Nezda et al. 2006). Arabic language is the native language of over three hundred million people in over 22 nations (Shaalan 2010), giving credence to the importance of creating NER frameworks for the Arabic text. Arabic language possesses several features and qualities

that make it challenging to develop an effective NER framework. This sub-section is dedicated to describing the challenges faced in Arabic language recognition.

a. Arabic script

Because the Arabic script is not confined to a single language family, it displays different linguistic properties and is used to write Arabic language as well as several others such as, Urdu, Kurdish, Persian, Pashto, Malay language referred to as Jawi, Old Turkish language, among others (Habash 2010). Little progress has been noted in Arabic-script-based languages computational processing following the first work conducted by Hlal (1985). However, several researchers have brought forward computational tools for Arabic languages and resources that are based on modified Arabic scripts as opposed to original Arabic scripts like lexical sources as illustrated by Halpern (2009), and Arabic Morphological analyzer in Buckwalter (2002) study.

Moreover, dialects and languages that are Arabic script-based like MSA and CA provide numerous homograph and word sense uncertainty in different degrees. The lack of representation of short vowels in typical messages significantly contributes to the ambiguities. Normally, the number of ambiguities of a token in numerous dialects is 2.3, but in MSA it is 19.2 (Buckwalter 2002). Issue management is an actual challenge in NLP of Arabic-scripted languages. This is exemplified by the determination of the NLP uncertainty concerning the representation of phonetic and logical information and the space and world learning. The ambiguity lies in the calculation of phonetic information used to address the issue. Arabic ambiguity is insurmountable at every lexical, morphological and syntactic level.

b. Capitalization issue

One of such challenges is issue of lack of capitalization. In Arabic orthography, there are no capital letters to differentiate initial proper names letters like other languages that are based on Latin-script (Elsebai et al. 2009; AbdelRahman et al. 2010; Attia et al. 2010; Oudah & Shaalan 2012; Aboaoga & Ab Aziz 2013; Shaalan 2014; Zaghouani 2014; Zirikly & Diab 2014; Elsayed & Elghazaly 2015; Alanazi 2017). This makes the

detection of NEs, expressed in one word or sequence of words quite difficult (Farber et al. 2008). The ambiguity rising from the lack of capital letters is compounded by the way most Arabic places, proper nouns or things (NEs) are not distinct from common nouns and descriptive words that are non NEs. As a consequence, a method that is dependent on nouns dictionaries employed to resolve this issue is uncertain (Algahtani 2012). For instance, the Arabic proper noun أكرم (Akram) relays different meanings in a sentence based on the context in that it can function as a verb (honored), or a person's name (Akram) or a superlative (the more generous).

Moreover, NE is frequently integrated in the sentence context through, for instance, the use of prefix/suffix or cue words at the beginning or the ending or the NE. Hence, this ambiguity is resolved through a more robust NE analysis as it imposes challenges in accurately identifying NE. Consider the sentence … that means 'the falling of his heart is in Sana'a, with the correct identification of the triggering constituting مسقط فؤاده is one expression that indicates that the place that the individual loves helps in identifying that the place is a noun.

c. Agglutination

Agglutination refers to a common occurrence in Arabic text as akin to other Latin-scripted languages. Different lexical variations can be extracted from different agglutination patterns. One word may be developed from one or more prefixes/suffixes and stem/root having different combinations, and this will generate complex and systematic morphology. Additionally, clitics are present, where in other languages, they are deemed to be distinct words, but in Arabic text they agglutinate to various words (Shaalan 2014; Alanazi 2017; Cotik et al. 2017; Najar & Mesfar 2017; LanguageComputer n.d).

In Arabic language, clitics are integrated into NEs, among which conjunctions like و (wao-and), ف (pha-if), and prepositions like ك (k, as, like) or a combination of the like, such as وف (wa-pha- and then). NER normally hinges on the NE words context that could occur in different morphological forms (Salah & binti Zakaria 2017). Moreover,

the issue of sparseness would call for robust training corpora and to steer clear of it, this needs morphological pre-processing. Some of the solutions that have been brought forward by prior studies (Grefenstette et al. 2005; Farber et al. 2008; Alkharashi 2009) have their basis on ignoring prefixes/suffixes and retaining the original morpheme.

This can be exemplified by processing of the word وبسوريا (and by Syria) that results in Syria as the location name. Other brought forward solutions include segmentation and separation of text and then delimiter between the relevant morphemes to retain the information contained within the sentence context (Benajiba & Rosso 2007).

d. Auxiliary vowels

There are some diacritics in Arabic language representing vowels used to change the meaning of a word after which a different meaning can be obtained. The NEs challenge stems from the fact that word short vowels are ignored in most texts in Arabic and they represent those that affect the phonetic representation (Smrz 2007).

However, majority of Arabic texts written in CA or MSA lack diacritics and this leads to single-to-multiple meaning ambiguity (Alkharashi 2009). Hence, inaccurate morphological analysis arise when applied on the same nominal words. Diacritics are also lacking in printed and digitized media sources like newspapers, magazines, blogs and the like as native Arabic speakers do not require them to understand the meanings of the words as their meanings is understood through the contexts. Such identification is still impossible through computational systems.

Furthermore, the lack of understanding of diacritics in major Arabic texts has led to the ambiguous meanings of lexical types representing various meanings. Hence, researchers have been attempting to resolve such issue via contextual information analysis and efficient comprehension of the language (Benajiba et al. 2009).

This can be explained by considering the word نور, which refers to the proper name (noor-light) or the verb (Nooar-enlighten), or a female name. The challenge

compounds if the word contextual information is also ambiguous owing to non-vocalization (Mesfar 2007). Another instance is the word حكم that may refer to two distinct proper names, which are (hakam-judge) or (Hakama-rule) and thus resulting in triggering words that could lead to different types of NEs.

e. Divergence in writing styles

The Arabic language possesses some transcriptional vagueness that is related with the NEs borrowed from other languages. The issue arises from the different transliterated ways a word has (Shaalan & Raza 2007) and such vagueness owes itself to the differences in the usage of transcription methods and the way in which the words are written by writers of Arabic (Shaalan & Raza 2007; Halpern 2009). Moreover, standards and guidelines have yet to be established to govern the writing styles of transliterated words, and hence, the transliteral word can be written in different spellings that could compound the ambiguity that computation systems are faced with.

For instance, the English word 'google' when transliterated into Arabic can be written in different spellings with the use of Arabic scripts, despite the sameness of its meaning - غوغل ، قوقل or جوجل and the reason behind the different styles in borrowed words is that Arabic is more speech-oriented compared to Latin-scripted languages. This may lead to different forms of writing for one NE. A method that was proposed to resolve such ambiguity is to confine the related transliteral words describing the same NE and to connect them to represent one NE, or by normalization as proposed by (Pouliquen et al. 2006; Refaat & Madkour 2009; Steinberger 2012).

f. Named Entities inherent ambiguity

Arabic computational systems, similar to other languages, are also challenged with the mutual ambiguity issue that concerns two or more NEs and this compounds the critical challenges in creating NLP systems for Arabic NEs (Attia 2008). This can be explained by the fact that some studies revealed 21 different analytical outcomes produced by BAMA for the Arabic word (ثمن) (Maamouri et al. 2009). In the case of Arabic NEs

nearby in a sentence, an inherent ambiguity exists that should be addressed – for instance, the sentence ابو موسى فاز بالمركز الأول (Abu Musa won the first position), where Abu Musa can refer to a person's name, or a name of a location and thus conflicting outcomes will be tagged as names of both location and person. In some studies, the authors have proposed approaches that can resolve this issue via suggesting cross-recognized NE (e.g., Shaalan and Raza (2009)), where one NE is selected. An alternative solution came from Benajiba et al. (2008), where the preferred NE is chosen, for which the optimum accuracy is reached by the classifier. Inherent ambiguity of NEs may also take the following forms;

i. Homographs

These refer to words in Arabic having the same spelling but possessing different meanings based on the word's context used in a sentence (Salah & binti Zakaria 2017). For instance, the word ذهب can mean NE noun (gold) or verb (went).

ii. Internal word structure ambiguity

When segmented in different patterns, some Arabic words can lead to different meanings and this can be explained by considering the word وفي, when segmented into و + ف + ي can mean (and in that) but when left as a whole, وفي, it means (faithful).

iii. Syntactic ambiguity

Under this type of ambiguity, the ambiguity stems from the syntactic status or NEs in the Arabic text (Farghaly & Shaalan 2009). For instance, رأيت رئيسة الجامعة الجديدة may translate to "I saw the new president of the university", or "I saw the president of the new university", depending on the syntactical status of different NEs in a sentence, which in this case is رئيسة (president) and جامعة (university).

iv. Semantic ambiguity

The phrases and sentences in an Arabic text can relay more than a single meaning and this can be explained by considering the sentence سالم يحترم امين اكثر من احمد (Salem Yahtaram Ameen Akthar min Ahmed – Salem respects Ameen more than Ahmed). The meaning can either be, Salem respects Ameen more than Ahmed does, or Salem has more respect to Ameen than to Ahmed.

v. Anaphoric ambiguity

The statement قال سامي انه تزوج literally means "Sami said that he got married", leading to the confusion in NEs recognition within the context as the sentence can be interpreted as if Sami is the one who got married or another person did.

vi. Compound named entities

Compound NEs are a combination of many words and in this case, it is difficult to recognize their beginning and their end. For instance, عبد الرحمن may indicate a single NE (a name of a person) or two NEs (Abd-servant and Alrahman- one of God's names meaning the Merciful).

vii. Ambiguity in acronyms

It is easier to distinguish English acronyms compared to Arabic ones as in the former, acronyms are written in capital letters (e.g., CNN). However, in Arabic acronyms NEs are difficult to recognize as they are not capitalized (e.g. سي ان ان).

g. Systematic Spelling mistakes

Spelling error involves writing the word in a way that contradicts with the Arabic spelling rules that can be because of the writer's ignorance, or error in typography or technical issues. The spelling and grammatical errors issue is common when writing in

Arabic. In fact, systematic errors in spellings are often made by Arabic writers, particularly when it related to distinct characters owing to their similar script and characteristics (Shaalan et al. 2012). Such mistakes can lead to confusion in the meaning of the word because of differences in orthography (El Kholy & Habash 2010; Habash 2010; Al-Jumaily et al. 2012).

For instance, ة (Taa-closed Taa) is an Arabic letter with similarity confusion in terms of character and it is normally used to denote morphological feminine NEs, as with the character ه (Haa-opened Taa). The problem occurs as the two characters are sometimes interchangeably written among Arabic writers (Farghaly & Shaalan 2009). Another character confusion is between the letters ظ and ض owing to the similar phonic characteristics. In this regard, the word ضل (Thal-get lost), can be misspelled as ظل (Thel-shadow) (Shaalan 2014).

h. Lack of resources

The effectiveness of Arabic NER system requires testing the different sets of tagged resources like corpora and gazetteers (collection of the places NEs) owing to their role as sources on which the NER system implementation and performance testing can be achieved. For the Arabic language, such resources are lacking and the few available ones have confined coverage (Abouenour et al. 2013). In addition, creating sufficient Arabic new resources would be too expensive (Huang et al. 2004; Bies et al. 2012) and as such, researchers are largely dependent on their own resources that still require further validation and enhancement (Bontcheva et al. 2017). In this case, some corpora developed by individual researchers can be used for public use (Benajiba et al. 2007), whereas others are kept confidential under license agreement (Strassel et al. 2003). Moreover, owing to the more current attention garnered by the Arabic language NER systems, it is not uncommon to see some Arabic corpora with significant size online but majority of them still have limited tools to support functions for researchers that are Arabic corpus-based.

i. Dilemma of normalization

Arabic text normalization presents a challenge among researchers in their development of Arabic NER systems (Farghaly & Shaalan 2009). This is because the meaning of the Arabic words hinges on the related diacritics and large collections of Arabic texts are missing diacritics as the word's exact meaning's extraction from its context. This is still impossible for computational systems. This can be exemplified by the first letter of the alphabet in Arabic script (ا - Alif), that can lead to a different meaning, with the addition of some diacritics. More specifically, hamza ء can be placed above or below إ – أ, Eee-Aaa and it can take the form of a curly hyphen (آ-Aaa). For instance, the word أمل (Amal) with hamza on top represents a female name but a curly hyphen on top آمل, represents (Aamal-hope). This issue is addressed by normalizing the input text – in this case – replacing the alif with hamza, with just alif alone (Larkey & Connell 2001) . However process of normalization could result in ambiguous outcomes in certain words; for instance, when hamza is removed, the following different Arabic words will remain the same; أن and إن (Farghaly 2010).

j. Arabic Diglossia (language variants)

Similar to other languages, Arabic has various forms that are currently in use, which are Classical Arabic, Modern Standard Arabic and Colloquial Arabic, with the latter related with certain group of people or region. For instance, Classical Arabic is the primary language form used in religious matters (during prayers), MSA is used in reading news, and lastly the colloquial form of Arabic is mainly utilized when talking with relatives and friends hailing from one specific region (Farghaly 2004).

Several challenges are faced during the development of NER systems in terms of diglossic language, like Arabic. The major problem is the impossibility of developing a single NER system/application that is geared towards different language variants as each type possesses distinct characteristics including different grammar, spelling and even morphology (Shaalan 2014; Salah & binti Zakaria 2017; Salah & binti Zakaria 2017).

Some researchers in literature like (Larkey & Connell 2001; Habash & Rambow 2005) dealt with Arabic diglossia using a method depending on two-layered approach, with the first stage involving the development of NLP for colloquial language. This is carried out through the extraction and classification of grammar and morphology of the colloquial language, transforming them as close and as similar to MSA as possible and using MSA NLP tools on them. This is similar to the study conducted by Shaalan et al. (2007), where the authors converted Egyptian Arabic text into MSA by adopting processes pertaining to lexical and transformation.

2.5 RELATED WORK ON ARABIC NAMED ENTITY CORPORA

The past decade has witnessed construction of the background information resources to overcome several challenges in text mining tasks. For non-English languages with poor knowledge sources such as Arabic (Bontcheva et al. 2017), these challenges have become more prominent especially for handling the NLP applications that require human annotations (Maamouri et al. 2004). In the NER task, several researches have been introduced to address the complexity of Arabic in terms of morphological and syntactical variations (Karaa & Slimani 2017). However, there is small number of studies dealing with Classical Arabic (CA), which is the official language of Quran and Hadith. CA was also used for archiving the Islamic topics that contain a lot of useful information which could be of great value if extracted.

Previous work on Arabic Named Entity corpora has been annotated either manually or automatically. There are some very useful resources for the named entity recognition task such as the early work Benajiba et al. (2007), in which he builds annotated Corpora called, ANERcorp, a manually annotated corpus in Arabic which is created to be used in Arabic NER tasks. It consists of two parts; training and testing. It has been annotated by one person in order to guarantee the coherence of the annotation. There are more than 150K tokens in the corpus and 11% of them are Named Entities. Every token in the corpus is annotated with one of the followings; person, location, organization, miscellaneous or other. The corpus is selected from news wire and other types of web sources. Benajiba et al. (2007) also built NERGazet which contains three types of gazetteers built manually (Person: 1950, Location: 2309, Organizations: 262).

On the other hand, there are several researches that have been introduced to exploit Wikipedia as a knowledge resource for ANER and classification. Wikipedia is a multilingual collaboratively constructed largest free encyclopedia containing semi-structured data. It contains concepts on a wide range of topics such as science, history, health, politics, and news events to contributions by collaborators. Recent studies have shown that Wikipedia is a reasonably accurate resources in many applications/tasks such as measuring semantic relatedness (Gabrilovich & Markovitch 2009; Saif et al. 2016), word sense disambiguation (Mihalcea 2007; Moro et al. 2014), building or enriching lexical sources (Navigli & Ponzetto 2012; Saif et al. 2015) and NER (Nothman et al. 2013; Saif et al. 2013).

Another works by Mohit et al. (2012), known as AQMAR Named Entity Corpus, is a 74,000-token corpus of 28 Arabic Wikipedia articles hand-annotated for named entities. The corpus focusses on four domains; Entity types in this data are POL categories person, organization, location.

Attia et al. (2010) proposed a method to automatically create a NE lexicon by exploiting Arabic WordNet and Arabic Wikipedia. This method consists of the following steps: mapping, NE identification, post-processing and discretization. To classify entities in the nodes of semantic taxonomy, the Lexical Mark-up Framework has been used for representing the entities. The extracted resource contains approximately 45,000 Arabic NEs and can be used with different levels of granularity for NE recognition. The evaluation of the lexicon achieves precision scores from 95.83% (with 66.13% recall) to 99.31% (with 61.45% recall) according to different values of a threshold.

Using Wikipedia as a Resource for ANER, Alotaibi and Lee (2012) described a supervised machine learning A Conditional Random Field (CRF) classifier to predict the presence of the named entities in the Arabic Wikipedia. The described method has been evaluated on a random sample of Wikipedia texts and achieves 88.62% F-measure of detecting both simple and complex named entity phrases.

Azab et al. (2013), compiled CMUQ-Arabic- NET Lexicon corpus, a lexicon of

about 57K named-entity pairs, an English-Arabic names dictionary from Wikipedia as well as parallel English-Arabic news corpora with four classes of NEs: Person (PER), Location (LOC), Organization (ORG) and Miscellaneous (MISC). They used off-the shelf NER system on the English side of the data.

Recently, the knowledge-based approach by and (Saif et al. 2013; Saif et al. 2015) 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 belonging to several categories, these categories can be exploited to recognize the different named entity types.

In short, most of the corpora of NER have been introduced to alleviate the issues that are related to Arabic NER. These corpora have been formatted using XML annotation standards to make them easily evaluated in the several tasks. However, these collections with the size ranging from 14k to 230k cover only named entities in modern standard Arabic such as person's names, some organisations and geographical locations names (Zaghouani 2014). There is also some automatic creation of Arabic named entity annotated corpus with small size in modern Arabic. Therefore, in this study, named entity corpus in classical Arabic that focuses on Islamic domain is created to satisfy the need for a new corpus.

2.6 RELATED WORK ON NAMED ENTITY RECOGNITION

The first task in the process of information extraction is NER. In this regard, Marrero et al. (2013) contended that the term for NE as referred to by majority of researchers and mentioned in conferences can be categorized into four namely, named entity as proper nouns (serving something/someone as a name), names used as rigid designators (a universal given concept), entities as distinct identifier (the concept referred to is unique within a specific concept), and lastly, named entities (based on the applications purpose and domain). Moreover, NEs have to be defined as mentioned to keep them consistent with literature, forums and tools that have been extensively utilized Marrero et al. (2013). For the purpose of the present study, the definition provided by [MUC-7],

which stated that NEs are unique identifiers of entities (individuals, locations, organizations) times (dates, times and durations) and quantities (money, measures, percent and cardinal numbers).

Therefore, NER refers to the detection and classification of NEs into categories. The categories are defined through the three leading bodies, which are MUC, CoNLL and ACE – the three dominating bodies of evaluation and definition of various tasks that are NER-related. This work chooses the NE tasks as defined by the MUC-7 conference, which are ENAMEX, TIMEX, AND NUMEX. First, ENAMEX represents people, organizations and location names, TIMEX represents time and date, and lastly, NUMEX represents monetary values and percentages (Chinchor & Robinson 1997). For the purpose of this study, the researcher confines the focus to ENAMEX, TIMEX, NUMEX, and also other types related to Islamic domain detection and classification.

NER is among the primarily elements of information extraction in the process of extracting various properties of unstructured text, but it is more than just an IE component. Its applicable to other NLP application have garnered the focus of researchers whose works are dedicated to Machine Translation, Speech Recognition, Information Retrieval, Question and Answering, among others (Oudah & Shaalan 2012; Alanazi 2017; Aljasser & Vitevitch 2017).

This section is dedicated to present the related works on the three approaches that have been used in this study.

2.6.1 Rule-Based Approach

A set of manually written and defined rules by linguists are used in this approach (Srivastava & Ghosh 2013). The systems comprise of a group of patterns that use grammar, syntax, and orthographic features combined with dictionaries (Mansouri et al. 2008). Systems with this type of approach are better performers compared to other approaches for domain specific tasks (Srivastava & Ghosh 2013). Nevertheless, the drawback of this approach is the tiresome creation of manual rules, the expense incurred and the enlisting of the entire rules needed to identify and classify NEs. However, upon

the completed definition of the entire required rules, the system can work efficiently in its detection and classification of NEs. Because the approach is language and domain specific, the system cannot port to another language/domain.

Majority of approaches under this type employ human-made rules for the extraction of NEs. These approaches generally consist of grammatical rules as their patterns and they leverage the handcrafted rules for the NER task. The approaches have their basis on grammar rules stemming from the linguistic knowledge and the list of names for the accurate detection of complex entities. Added to this, rule-based systems are composed in the form of finite state transducers or regular expressions as illustrated by Grishman (1997). The followings are the important rules and techniques that can be used in Arabic rule-based NER systems:

a. List Look-up Techniques

This technique hinges on lists that can exist in different forms (gazetteers, white list and dictionaries) that can be obtained based on some corpora of language (Shaalan 2014).

i. Corpus

Corpus is useful in a large collection of annotated texts, where the identification of NEs is based on their types. Corpus can take a specific or general form or it can cover one domain (politics, economics). Because of the monumental research progress in Arabic language, several corpora have been proposed, like NooJ (Mesfar 2007), ACE and Treebank Arabic datasets (Shaalan & Raza 2008), arabiCorpus (Traboulsi 2009), and ANERcorp (Benajiba et al. 2007).

ii. Gazetteer

A gazetteer refers to a list of defined NEs and it comprises a particular name list for specific NEs category (Elsebai et al. 2009; Shaalan 2014). This can be explained by the fact that Malaysia is in location category of gazetteer, whereas Sami is the category of person names. Gazetteers term is interchangeable with whitelist and dictionaries

(Shaalan & Raza 2008). Whitelist primarily has fixed strings of texts that are taken as NEs without further identification mechanisms used (e.g., application of grammar rules. Moreover, the whitelist entries may take the form of one word or multi-word expression like محمد حامد الغزالي (Mohammad Hammed Algazali), but in the dictionary, they may compose of an NE-related name, or place that is not related to NE; for instance, أحلام (Ahlam) may be a person's name or it may convey the meaning of 'dreams'.

iii. Blacklist (filter)

Blacklists, sometimes referred to as filters are utilized to reject words or words strings that are considered as invalid NEs; for instance, وزير الشؤون المالية المدير العام (the financial affairs minister and the general manager), where the phrase المدير العام (the general manager), is not a valid NE. The blacklist primarily rejects such NEs and they could take the form of words or phrases. This makes for accurate NER system as they are used as filters that sieve ambiguous expressions that cannot be accurately detected through the use of system rules and dictionaries (Shaalan & Raza 2009).

iv. Stop-words

Stop-words list generally has a group of words that are NE related but they are not valid NEs and hence, should not be a part of NEs. For instance, Arabic prepositions, specifically, انتخب في الدورة الأولى (elected in the first round), with the word في (fee-in) belonging to stop-word list as mentioned by Elsebai et al. (2009).

v. Trigger words (keywords):

Trigger words, often referred to as keywords can be described as frequent words surrounding NEs – normally prior to or after NEs. They help in identifying NEs and can be in the form of verb list or noun list –for instance, قال (said). These words are utilized for the identification of NEs in short phrases (Shaalan & Raza 2007; Elsebai et al. 2009; Aboaoga & Ab Aziz 2013).

b. Linguistic techniques

In the context of rule-based systems of Arabic language, the linguistic methods depend on the rules and pattern of Arabic language writing used for the extraction and recognition of NEs (Salah & binti Zakaria 2017). The performance of such systems depends on the rules robustness in identifying different NEs types and some of the rules are presented as follows:

i. Grammar Rules

Grammar rules are the set of grammatical rules that are used for the recognition of NEs (Mesfar 2007; Traboulsi 2009). They help in forming well-structured sentences and as such, their application in NER needs the knowledge of the processed language. Additionally, the grammar rules identify NEs from text through the use of pre-established rules and patterns. Because of the complexity of grammatical structure of Arabic language, particularly Classical Arabic, the grammar rules are important in recognizing NEs. Taking the instance of a sentence الملك السعودي عبدالله (the Saudi King Abdullah), the used suffix honorific word is الملك (the king), and the second indicating word is السعودي (Saudi), which is nationality, leading to the identification of the word عبد الله, as the NE.

ii. Heuristics Rules

Heuristic rules are general rules that are dependent on the type of the overall rule-based approach employed in the NER system as exemplified by their use to enhance the capability of the applied trigger words in the system (e.g., NEs trigger words) (Elsebai et al. 2009).

iii. Morphological Rules

Morphological rules are produced on the basis of the normal construction of words from stems, prefixes and/or suffixes. Majority of the Arabic words in a sentence generally contain the word's stem and other suffix and/or prefix (Farghaly & Shaalan

2009). For instance, the word يذهب (yadh-hab) is a combination of ي+ذهب with the stem ذهب and prefix ي While grammar rules operate on the sentence structure for their identification of NEs, morphological ones deal with the words structure to do the same – they specifically examine the word stem, affixes and/or suffixes. A popular approach that researchers often use for NER using morphological rules is Buckwalter Arabic Morphological Analyser (BAMA) (Buckwalter 2002). BAMA primarily operates different tables, including a table to collect Arabic stems, prefixes and suffixes; for instance, بمكه (be-Makkah), ب is considered by the morphological rule as a prefix, and thus removes it from the word and the remaining root is مكة (Makkah), referring to location NE.

More importantly, TAGARAB is one of the pioneering works in Arabic NER (Maloney & Niv 1998) and it incorporated supporting data and a pattern-matching engine having different components of morphological analysis in order to determine the five categories of NEs, which are person, location, organization, time and number. It had a recall of 80.8%, precision of 89.5% and F-measure of 85%, using randomly chosen datasets taken from Al-Hayat. This method was used in text encoded in ISO-8859-6 that was originally exposed to the tokenizer. As a result, the tokenized stream was employed by the Name Finder Module. Such module comprises of two units, which are morphological tokenizer and named finder. In the system, stem lists and feature extraction rules are employed as the inputs to finite-state scanner, whereas word lists and pattern-action rules are the inputs to NetOwl Turbo Tag™ pattern engine. The final result comes in the form of annotated text with suitable SGML tags for every item that is extracted.

In a related study, Mesfar (2007) created and proposed an Arabic NER system using NooJ linguistic platform. The system components included a gazetteer, tokenizer, triggers and morphological analyzer to recognize proper names, dates and temporal expressions utilized in Arabic text. The system was used to evaluate part of the Arabic version of Le Monde Diplomatique and the findings depended on the types of individual NE – precision, recall and f-measure. The figures differed between 82%, 71% and 76% for names of place, 97%, 95% and 96% for numerical expressions coupled with time.

The average accuracy of F-measure was 87% and in this approach, the standard Arabic text is inputted into Nooj tokenizer that outputs it in text form, after which the morphological analyzer process inputs from text forms, lexicons of simple inflected forms, and morphological grammars to produce recognized forms related with linguistic information. Moreover, the approach entails using the gazetteers and syntactic grammars for the production of NEs.

Similarly, in Shaalan and Raza (2007) study, the authors developed a system referred to as Person Named Entity Recognition for Arabic (PERA). The system made use of linguistic grammar-based techniques with whitelist dictionaries to recognize person NEs in Arabic text significantly and accurately. There are three components to PERA and they are, name lists called gazetteer, regular grammatical expressions forming the lexicon, and filtration mechanism via the established grammatical rules that assist the exclusion of invalid names. Both ACE and Treebank was also employed, coupled with internet sources. The results indicated that PERA achieved the following precision of 85.5%, recall of 89% and F-measure of 87.5%.

The above system was further enhanced by the authors (Shaalan & Raza 2008; Shaalan & Raza 2009) in the form of Named Entity Recognition for Arabic (NERA). NERA is described as a hand-crafted rule-based system that has three components (dictionaries, grammar rules of regular expressions, and filter mechanism). The same method and functionality was used by the authors in NERA as they had in PERA, the only difference being that the former has the capability of supporting ten types of NEs (locations, persons and organizations, price, data, ISBN, time, phone number, file names, and measurements). In the process of evaluation, the authors made use of resources from ACE, the Web newspapers, the Quran and Arabic literature, in order to develop distinct corpora, and for in-depth extraction of semantic information. NERA was successful, with the following percentages in persons, with F-measures of 87.7%, in locations with F-measures of 85.9%, and in organization with F-measures of 83.15%. The system also exceeded 90% performance, an average for the entire MUC NEs.

Along the same line of study, Traboulsi (2009) proposed an NER system that he developed through the use of local grammars. The system managed to identify

consistent structures of person names occurring often in the news text by using several sources for the corpus (arabiCorpus), collected from the newspapers archive (i.e., Al-Hayat, Al-Ahram, Al-Watan issued in Kuwait and At-Tajdid, issued in Morocco). The system employed some corpus extracted from the Holy Quran, Arabic novels (1001 Nights and medieval medical and philosophical works). The address expressions, time and date were extracted from letters through the use of the above corpus. His method depended on the use of corpus linguistics, methods and techniques as a result of which, it is consistent with the local grammar formalism to identify patterns related to person names in Arabic news texts.

Furthermore, , Al-Shalabi et al. (2009) developed an Arabic NER algorithm with the objective of retrieving Arabic proper nouns by using lexical triggers. The authors focused on regional patterns of consideration like name connectors. The algorithm was developed to recognize seven types of NE (person names, major cities, locations, countries, organizations, political parties and terrorist groups). But the reported research only stressed only on the person. Added to this pre-processing of input for data erasing and removal of affixes depended on heuristic rules used by algorithm. As a consequence, internal evidence triggers like connectors of person name were utilized for the identification of NEs. Aside from the above, the system assessed the use of a total of 20 documents that were randomly chosen from Al-Raya newspaper and it managed to reach an overall precision of 86.1%.

Added to the above studies, Elsebai et al. (2009) examined grammar rules in Arabic text and developed a set of keywords for person name. The system used pattern-matching with Morphological Analyser and it notably indicated optimum performance based on F-measure of 89% that exceeded that of PERA.

Zaghouani (2012) employed rule-based approach on RENAR system divided into three levels (morphological pre-processing, known name identification and applying local grammar). The system aimed to identify named unknown named entities. RENAR calculated precision, recall and F-measure equated to 87.17%, 65.74% and 74.95% respectively, with ANERsys of 1.0, ANERsys of 2.0 for person, location and organization, after the application of ANERcorp dataset. The system's performance was

then to be 73.39% for precision, 62.13% for recall and 67.13% for F-measure. The method had its basis on three processes namely pre-processing, lookup of full known names and recognition of unknown names through local grammas and a set of dictionaries. Names caught repetitively (at least twice) in long-term multilingual news, were checked in a manual way and for name retention, they were kept in a database.

Meanwhile, Asharef et al. (2012) concentrated on the domain of crime and built small corpus for this purpose. The system made use of rule-based approach for Arabic NER that was developed using syntactical rules and patterns, by taking features like prefix/suffice of current word, morphological and POS information, surrounding words information, and tags into consideration. The system also took predefined crime and general indicator lists and Arabic named entity annotation corpus taken from the domain of crime. The system showed an overall performance of 91% (precision), 89% (recall) and 89.46% (F-Measure).

An Arabic NER system was also built by Aboaoga and Ab Aziz (2013) in their study, with the aim towards recognizing names of persons. Their system depended on trigger words to identify the person in various domains –for instance, مدير (manager), رئيس (president), عميد (dean), لاعب (player), حكم (referee), مدرب (coach), مشرف (supervisor), and معلم/مدرس (teacher). They gathered corpora from the archives like online Arabic newspapers (koora.net, aleqt.net, and Alquds.net) and used sentence splitter and tokenization with gazetteers on the domain of politics, economics and sports. The latter domain displayed better performance compared to the former two. The average F-measures for person names recognition achieved 92.66%, 92.04% and 90.43% in sports, economics, and politics, respectively.

In a more current study, Elsayed and Elghazaly (2015) proposed an NER system to upgrade the recognition of NEs, specifically for Arabic nouns. The system extraction method was based on two-sided approach – Arabic morphology and grammar, both coupled with gazetteers. The system was successful in identifying names of persons, titles, cities, countries, nationalities, dates and times in MSA, achieving an F-measure of 84%. The system employed Essex Arabic Summaries Corpus (EASC corpus) as the

dataset and the two types of rule-based approach namely, linguistic and list look-up methods, with the former using Arabic morphology and grammar rules sans gazetteers, and the latter using gazetteers. The gazetteers used a part of the GATE system developed based on lists of titles, names of persons, countries and cities. Moreover, the NE nationalities were obtained from the list of the countries.

Moving on to the knowledge-based approach (Saif et al. 2013; Saif et al. 2015), brought forward a system to classify concepts in the linguistic resource into NEs and linguistic terms. The authors employed Wikipedia as a semi-structured resource to determine the NEs (person, organization, location, events and media). Because every Wikipedia written article is linked to several categories, the categories can be leveraged for the recognition of the different types of entities. The trigger words were used to identify the NE types in short phrases. For instance, the keywords of an NE person can have terms like أشخاص (people) that can in turn be utilized to form patterns *_أشخاص, indicating that the people's categories أشخاص_على_قيد_الحياة (living people) and أشخاص_من_أصفهان (people from Isfahan). Bill Gates was assigned to living people, American billionaires, American technology writers and American investors. The triggers words of people are billionaires, writers and investors, classified under NE person. The trigger words for location include terms that are linked to places (e.g., cities, countries, village, rivers and capitals) and for NE of organization, it includes terms linked to the organization (e.g., companies, corporation, association, union and institution. Lastly, for events, the trigger words are linked to terms of events (e.g., wars, matches, championships, revaluations, elections, festivals, parties and invasions). Hence, the concepts of World War II, Japanese invasion of Manchurian, Mukden Incident, Berlin Blockade, Aden Emergency and Yemen's revolution was categorized under events.

Related work also came from OUDAH and SHAALAN (2016) that developed a rule-based NER system to enhance their performance and enable updating through automated rule. The mechanisms managed to develop the ability of recognizing decision conducted by the NER hybrid system for the determination of the rule-based component drawbacks. This, in turn, worked towards deriving new linguistic rules that

had a tendency to improve the rule-based system. Based on the empirical findings, the enactment of the enhanced rule-based system (NERA 2.0) creates the coverage of mis-categorized names (persons, locations, and organizations) with the following percentages, 69.93%, 57.09%, and 54.28% respectively. The summarized version of the studies in the literature review for rule-based system for Arabic language is displayed in Table 2.3.

Table 2.3 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 Tem-poral 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 news paper	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. (2015)	Wikipedia	Person, Location, Organization, Events and Media (movies name, songs, video clips, series)	General MSA	NEs were used for enhancing mapping technique

From the above table (Table 2.3), it is evident that majority of the studies in literature employed rule-based approach focused on MSA, as MSA is the current type of Arabic utilized extensively. Classical Arabic on the other hand, was not extensively used by researchers although this could open avenues for new research direction owing to its relationship with Islamic religious texts. Additionally, owing to the lack of resources in Arabic language, majority of the works made use of the same corpus as evidenced by the summarized table. This has limited the research focus on just a few domains including politics and economics, with the exclusion of medicine and religion, among others. With regards to the methods used in the rule-bases system, majority of them adopted list look-up approach using gazetteers and dictionaries and only less adopted Arabic language rules. This is because Arabic grammar is complex and thus, more efforts are required to eradicate the barrier through more and more studies using new rules of the language. The outcome of the evaluation of knowledge-based approach indicated that the rule-based system methods are capable of functioning efficiently and enhancing the natural language methods, including the measurement of semantic compositionality (Saif et al. 2013), and mapping of lexical sources (Saif et al. 2015).

To conclude, the development of studies dedicated to rule-based NER systems have been good so far but more efforts are needed to propose rule-based systems on MSA and CA, as the latter received the least attention. CA is related to the domain of religion and ancient Arabic literature (poetry, drama and novels) that constitutes a new avenue of research for Arabic NER. Therefore, there is a dire need to develop novel handcrafted rules that can employ grammars involved with Arabic language for the performance enhancement. Also, studies dedicated to the rule-based NER for MSA texts is still confined to only a few NEs kinds and limited domains. Further research is required to propose a new rule-based NER system that can propose NER in new domains (e.g., crime, sports, religion).

2.6.2 Machine- Learning Approach

The most extensively utilized NER approach in the Arabic language and other languages is machine learning (ML). ML methods employ features of text and words for NE recognition. The next sections provide summaries of the common features

employed in Arabic ML NER systems and related studies on such systems (Shaalan 2014; Salah & binti Zakaria 2017)

a. Features in NER used ML Systems

There are features in NER using ML, in the form of properties/descriptor attributes of words that are used for the recognition of NEs. In the NER for classifiers, the top essential task is feature engineering (Salah & binti Zakaria 2017). In this regard, word feature may be features specified through different ways through the use of various ways that use one or more than one Boolean/binary values, numerical/nominal values, with the following common features;

i. Word

Words refer to the distribution every NE type in the Corpus (AbdelRahman et al. 2010; Meselhi et al. 2014; Alsayadi & ElKorany 2016).

ii. Word-Left/Right

This pertains to the analysis of neighbor words or called surrounding words (left-right) of ‘n’ length. Analyses used are of numerous types using features such as part of speech, named entity tags from NES system (Abdallah et al. 2012; Shaalan & Oudah 2014; Zirikly & Diab 2014).

iii. Word Length

This feature is useful in checking if the length of a word is lower than three as very short words are not considered as entities (Abdallah et al. 2012; Shaalan & Oudah 2014).

iv. Special Marker

This feature is invaluable to identify the existence of special symbols/markets in the text (Oudah 2012).

v. Word Prefix/ Suffix

This feature utilizes pattern matching to identify word prefix/suffix of ‘n’ length (Abdul-Hamid & Darwish 2010), and they rarely come as NE - the feature could accurately indicate the existence of NE. Example of prefix and suffix are show in Table 2.4.

Table 2.4 Example of prefix and suffix

Word	Translation	Lemma	Prefix	Suffix
وعلمه	and taught him	/ taughtعلم	/ and و	him /ه
العربية	Arabic	/ Arabsعرب	/ the ال	ية /
رمزي	Ramzi	Ramzi	-	-

vi. Capitalization

This is a binary feature that indicates the presence of capitalization information on the gloss that corresponds to the Arabic word (AbdelRahman et al. 2010).

vii. Lexical match between Arabic and English

The Arabic-English lexical match can be realized through the use of bilingual lexicon of morphological analyzer (Benajiba et al. 2008). A good requirement for this may be exemplified by Google – a word that may be transliterated to Arabic as جوجل or غوغل or قوقل. In the training corpus, if Google has occurred in the first transliteration, the classifier is not able to classify the second one. Because of the multiple coiners for Arabic throughout Arabic countries, majority of untranslatable words have been relayed in many forms (Saif et al. 2015) to Arabic. For instance, ‘biome’ and ‘pixel’ have been transliterated to Arabic in different lemmas ‘بيوم’ versus ‘مجال حيوي بيئي and ‘نقطة ضوئية’ versus ‘عنصورة’ as found in WordNet and Arabic Wikipedia. In this sense, the bilingual features can lead to enhanced accuracy of the semantic similarity between the two concepts obtained from different resources of knowledge (Benajiba et al. 2008; Meselhi et al. 2014; Zirikly & Diab 2014; Alanazi 2017).

viii. Nationality feature

Under this feature, the combination between two types of lexical and contextual feature takes place; for instance, وصل الرئيس المصري محمد مرسي إلى تركيا (Egyptian President Mohammed Morsi arrived in Turkey). This feature is a binary feature that determines if the word is stored in the list of nationalities (Benajiba et al. 2008; Alshaikhdeeb & Ahmad 2016).

ix. Trigger words (key words) feature

A significant feature that can guide the identification of NE and is capable of adopting different forms like verb list or noun list is called an indicator feature. Under this feature, the word is identified as included in one of the lexical triggers lists or not. Several Arabic terms that have been examined to identify the NEs in natural language documents (Saif et al. 2013) have been referred to as Arabic trigger words to identify NEs in Wikipedia articles. Such trigger words are successful in classifying Arabic Wikipedia concepts employing category-based methods (Alsayadi & ElKorany 2016). Example of trigger words is show in Table 2.5.

Table 2.5 Example of trigger words

Type of NE	**Trigger**	**Translation**	**NE**
Person	قال	*said*	أحمد/ Ahmed
Location	سافر إلى	*Travel to*	دبي/ Dubai
Organization	شركة	*company*	سبأ/ Saba

x. Blacklist feature

This feature uses Blacklist dictionaries that contain entries that have to be rejected as NEs. This feature works in a two-fold approach to determine if the word is blacklisted.

For instance, in رئيس الجمهورية القائد الأعلى (the President supreme commander), the القائد الأعلى (The supreme commander) is deemed as an invalid NE.

xi. Stop words feature

This feature addresses frequent words that cannot be part of NEs and it works to determine if the word is in the stop words list (Benajiba et al. 2008; Alanazi 2017). Some of stop words are shows in Table 2.6.

Table 2.6 Example of stop words

Categories	The Word	Translation
demonstrative nouns	هذا	this
relative pronoun	الذي	who, which
adverbs	هناك	there

xii. Gazetteer feature

Under this feature, there are lists of specific information storage like people's names, organization names, location names, and days of the week, among others. It determines if the targeted words are present in the gazetteer category (Benajiba et al. 2008; Oudah 2012; Meselhi et al. 2014; Zirikly & Diab 2014; Atwan et al. 2016; Alanazi 2017).

xiii. Rule-based features

These are contextual features that cover NE types, with NE tags predicted with the help of rule-based NER system, deemed as features (AbdelRahman et al. 2010; Oudah & Shaalan 2012).

xiv. Infrequent Word

This feature is obtained by calculating the frequency of the word in the used corpus throughout the training phase and then choosing the cut-off frequency to develop the binary feature (Meselhi et al. 2014).

xv. Part-Of-Speech (POS) feature

POS is among the top important features often employed with ML and it identifies the word part of speech class, in the form of verbs, nouns pronouns, among others

(AbdelRahman et al. 2010; Abdallah et al. 2012; Oudah 2012; Boujelben et al. 2014; Meselhi et al. 2014; Zirikly & Diab 2014; Alsayadi & ElKorany 2016; Alanazi 2017).

xvi. Syntactic-based features

This feature uses the syntactic rules to label phrases that can take the form of noun/verb phrases (Alsayadi & ElKorany 2016).

xvii. Morphology-based feature

This constitutes a set of features obtained from the language morphology and it is extensively utilized. It is well-known for the production of morphology features in MADA (Farber et al. 2008; Habash et al. 2009; Meselhi et al. 2014; Shaalan & Oudah 2014; Alsayadi & ElKorany 2016; Alanazi 2017), with over 13 features, this come with more details in Chapter 7.

b. Learning Methods

Machine learning methods have more capability in comparison to rule-based methods owing to the fact that the system is trainable and it is suitable for different domains. The NER method transforms the identification issue into a classification one, after which statistical models are used to address the classification issue. The ML system acknowledges and categorizes NEs into specific classes like locations, persons, organization and others (Mansouri et al. 2008). Majority of current NE studies for the major languages, with the inclusion of Arabic made use of ML, also known as statistical approach. Moreover, ML algorithms have been extensively utilized for the determination of NE tagging decisions from annotated texts (Alanazi 2017; Alsayadi&Elkorany 2016). The approach to analyzing language n ML begins following a bottom-up strategy while searching for patterns/relationships to form. ML is of three types, which are supervised learning, unsupervised learning and semi-supervised learning. The most extensively published ML approaches for NER are Supervised Learning (SL) techniques that view the NER problem one that requires classification and the availability of significant number of annotated datasets. The common models

employed in ANER ML systems are;

i. Supervised approach

This approach is a pioneering applied method in ML systems and it aims to train data on specific patterns to determine it in the test part. It is a useful approach in the sentiment analysis field to train the data of the pattern that may relay positive/negative opinion (Altawaier & Tiun 2016). In this approach, considerable annotated corpora is needed with the statistical models for NER, with majority of the studies using several techniques including, Conditional Random Fields (CRF), Hidden Markov Model (HMM), Decision Trees (DT), Maximum Entropy Models (ME), Support Vector Machines (SVM) and Artificial Neural Network (ANN). These supervised techniques were used in ANER and some of the studies are presented and discussed below;

Conditional Random Field (CRF)

CRF refers to a statistical model utilized to segment and label data sequentially (McCallum & Li 2003). The model entails the use of several random and related features for the identification of NEs. CRF, according to (Lafferty et al. 2001) Laffer ty et al. (2001) as a probabilistic framework employed to segment and label the sequential data. Moreover, the CRF model can be calculated with the use of the equation below(Lafferty et al. 2001; McCallum & Li 2003) (1).

$$P(y|x) = \frac{1}{Z(x)} * \exp(\sum_t \sum_k \lambda_k \, f_k(Y_{t-1}, Y_t, x)) \quad (1)$$

In the above equation y stands for the sequence of labels, x depicts the sequence of data points, λ_k depicts the weights of every feature in the feature set, and $Z(x)$ depicts the normalization function. The equation shows that CRF may be deemed as an extension of ME and HMM. The efficiency of CRV has been validated in NER systems. Added to this, the CRF++ refers to a tool employed for the performance of CRF method, and it has been extensively used in various NER systems of Arabic language and other languages as well.

To begin with, Benajiba and Rosso (2008) conducted a study using the CRF method to replace Maximum Entropy to enhance the performance of the system. The system features include POS tags and Base Phrase Chunks (BPC), gazetteers and nationality. The findings revealed high system accuracy with the indicators of recall, precision and F-measure being 72.77%, 86.90% and 79.21% respectively.

In a related work, Abdul-Hamid and Darwish (2010) developed an NER Arabic system based on CRF for the recognition of three NE types namely person, location and organization. The system only focuses on surface features to the exclusion of other features. They tested the system using ANERcorp and ACE2005 dataset. The performance indicators on ANERcorp were found to be 89%, 74% and 81% for precision, recall and F-measure respectively. These findings confirm the system's accuracy more than what was reported in Benajiba and Rosso (2008) study.

Added to the above studies, AbdelRahman et al. (2010) combined two ML systems to address Arabic NER pattern recognition with the help of CRF and bootstrapping. The considered features were word-level features POS tag, BPC, gazetteers and morphological features. The system successfully managed to identify different NEs including person, location, organization, device, care, cell phone, date and time. The F measures of the NEs types were found to be 74.06%, 89.09%, 75.01%, 69.47%, 77.52%, 80.95%, 80.63%, 98.52%, 76.99% and 96.05% respectively. The results also confirmed that the system outperformed Ling Pipe NE recognizer when both were used on ANERcorp dataset.

Similarly, in Bidhend et al. (2012) study, the authors proposed a CRF-based NER system that is referred to as Noor, to extract the names of persons from religious sources. Ancient religious text corpora called NoorCorp were developed, with specific focus on 3 corpora, based on three Islamic books and jurisprudence in Arabic languages. The study also developed Noor-Gazette, to gazette religious names. The overall system performance of the new history, Hadith and jurisprudence corpora in terms of F-measure was found to be 99.93%, 93.86% and 75.86% respectively.

In the same study caliber, Morsi and Rafea (2013) examined the impact of

various features on the performance of conditional random field-based Arabic NER for Modern Standard Arabic test. The system employed CRF-based models and the authors developed baseline model to compare the results. The dataset was obtained from ANERcorp, and the system extracted four entity types, which were person, location, organization and others. The highest result of the system was noted to be 68.05 (F-measure).

Meanwhile, in another related study, Zirikly and Diab (2014) brought forward dialectal Arabic NER system with the use of Egyptian colloquial Arabic. The machine-learning approach made use of CRF approach for the recognition of NEs persons and locations and the NER features used are lexical with contextual features, gazetteers, distance from specific keywords and Broun clustering. The authors also developed an annotated dataset for the Egyptian dialect by manually annotating a part of the dialectal Arabic (DA) gathered from the linguistic data consortium (LDC) from web blogs. Furthermore, the annotated data was selected from a group of web blogs manually pinpointed by LDC as Egyptian dialect. The results showed that F-measure for locations and person names were 91.429% and 49.18% respectively.

In a more recent study, Zirikly and Diab (2015) developed an NER system without gazetteers for Social media, which is characterized by the use of both MSA and Dialectal Arabic (DA) through the use of supervised Machine Learning approach and CRF classifier. The results showed that yields an F1 score of 72.68%.

Alsayadi and ElKorany (2016) proposed an integrated semantic-based ML model for ANER. The author used a combination of several linguistic features and to utilize syntactic dependencies to infer semantic relations between named entities. The study extracted person, organization and location using CRF classifier. Results show that this approach can achieve an overall F-measure around 87.86% and 84.72% for ANERCorp and ALTEC datasets respectively.

Hidden Markov Model (HMM)

HMM refers to a statistical model using Markov process, possessing hidden states and

its mathematical equation was developed by Bikel et al. (1997) as shown in equation(2).

$$\log P\ (T^n|W^n) = \log P(T^n) + \log \frac{P(T^n,W^n)}{P(T^n).P(W^n)} \quad (2)$$

where T optimal tag sequence T = t1,t2,t3, . . . ,tn and W for a given word sequence W = w1, w2, w3 . . . , wn. The T in the above equation depicts the possibility of producing output of the HMM classifier. HMM has been used in the NER field, wherein the sequence of input mimics the sequence of dataset words, while the output tag sequence mimics the sequence of predicted NE classes for the words appearing in the sequence of inputs.

Lastly, an Arabic NER system was proposed by Dahan et al. (2015) that had its basis on HMM, using stemming process with the objective of addressing inflection and ambiguity in Arabic language. The system is works in complete automation while it recognizes Arabic person, organization and location NEs. The authors tested the system with the help of a developed corpus from France Press Agency, *Assabah* newspaper and Al-Hayat newspaper. The indicators of system performance were found to be 73% for precision and 77% for recall, while the F-measure found were 79%, 67% and 78% for persons, organization and location NEs respectively.

Decision Tree (DT)

The development of DT was attributed to the works of Sekine et al. (1998), where DT is a tree-like model that makes decisions at the tree nodes. The tree path depicts the sequence of decisions followed in order to acquire the terminal output represented by the leaves of the tree.

The tree's internal nodes depicted the features in the feature set, whereas the leaf nodes depicted the classes, and the branches depicted the feature values, indicating that the classification is conducted via traversing the tree. The DT construction is done by selecting the suitable feature at every internal node, producing offspring nodes for every features value, then splitting the data points over the children, and repeating the prior steps for every child. Feature selection of each node is done through information gain

(IG) as similar to the ones used in ID3 and C4.5 decision trees algorithms. The Information Gain formulate is presented with the following equation (Eid et al. 2011) (3).

$$IG(Y|X) = H(Y) - H(Y|X) \quad (3)$$

in the above equation, Y depicts the class, X depicts the feature and H depicts the entropy. The entropy is calculated by using the equation (Eid et al. 2011) (4).

$$H(Y) = -\sum_{i=1}^{k} P(y_i) \log_2(P(y_i)) \quad (4)$$

the feature having the highest IG is selected to be depicted by the internal node. Decision trees have been generally employed a classification method in varying NLP systems (e.g., NER systems).

In a study, Al-Shoukry and Omar (2015) used the ANER ML system with DT on the domain of crime in MSA. Their system was capable of extracting NEs of persons, locations, types of crimes, locations, times and date via DTC (Decision Tree Classifier) having extraction features. They obtained the dataset from online resources and they found the highest F-measure to be 81.35%.

Maximum Entropy (ME)

The ME model is used to predict the probabilities through the use of the least number of assumptions that differ from the applied limitations. Such limitations are obtained from the training data that expresses the features-outcomes relationship (Borthwick et al. 1998) with the help of the formula (Della Pietra et al. 1997) (5).

$$P(o|h) = \frac{1}{z(h)} \prod_{j=1}^{k} x_i \, f_i \, (h, o) \quad (5)$$

in the above formula, a_i depicts the weight of featuref_i, o depicts the outcome, h the history and $z(h)$ depicts the normalization function. Meanwhile, the parameters x_i are estimated through the Generalized Iterative Scaling (GIS) procedure McCallum

et al. (2000). The outcome having the highest probability is deemed as the predicted outcome (class) of an element (word). The weights of ME model are calculated through a tool called YASMET – a tool that has been widely used in several ANER systems adopting ME as the statistical method (Oudah 2012).

An ME Arabic NER system was brought forward by Benajiba and Rosso (2007), who developed an ANER system (ANERsys 1.0) using ME. The authors developed their linguistics resource called ANERcorp (annotated corpus) and ANERgazet (gazetteers). The used features are primarily contextual, lexical combined with gazetteers features and the system was capable of recognizing different NEs types (person, location and organization). However, the ANERsys1.0 system had issues with searching for NEs having compound structure that is made up of more than one word. This is why the authors came up with ANERsys 2.0 that has two level mechanism used to; 1) identify the beginning and end points of every NE and 2) categorize and identify NEs. The performance of the system in light of precision, recall and F-measure was found to be 70.24%, 62.08% and 65.91% respectively.

Support Vector Machine (SVM)

According to Cortes and Vapnik (1995), SVM is a popular technique in ML that is sometimes referred to as support vector network. It is supervised learning method involving other learning methods that analyze data for the purpose of classification and analysis.

Moreover, SVM involves the estimation of a hyperplane that categorizes the space elements into two groups +class and class-. The margin between the hyperplane and the nearest element requires maximization. It is estimated during the training phase by observing the features of every element in the training space with the real class. Every element is depicted by a vector of features and its actual class. The element's position juxtaposed to the hyperplane is what decides its predicted class by the SVM model. For instance, if the element is on the side of the + hyperplane then the predicted class is +. More specifically, the position is calculated by the equation (Vapnik 1995) (6).

$$g(x) = \sum_i^n w_i k(x, sv_i) + b \quad (6)$$

from the above equation sv_i depicts support vectors that are the closest in proximal data elements to the hyperplane, while N is the number of support vectors, $k(x, sv_i)$ depict the kernels for features mapping, w_i depicts the weights of the entire features of the element, and b is a constant that is calculated in the training phase. Because of its robustness to noise and its capability of dealing with considerable features sets in an effective manner, SVM has been extensively utilized in NER systems. Also, YamCha is a toolkit used to train SVM models and it has also been employed in various Arabic NER systems that adopted SVM as a statistical method.

The use of ANER with SVM was first attributed to the works of Benajiba et al. (2008), where the used features are contextual, lexical, morphological, gazetteers, POS tags and BPC, nationality and the corresponding English capitalization. The system has been evaluated and validated through the use of ACE Corpora and ANERcorp. The optimum outcomes were achieved when the entire features are focused on. Moreover, Benajiba et al. (2008) examined various NEs in light of their sensitivity to various features types and ACE datasets using SVM classifier. The optimum outcomes of F-measure were found to be 82.71% (ACE 2003), 76.43% (ACE 2004), and 81.47% (ACE 2005).

Further works from Benajiba et al. (2008) concerned the development of multiple classifiers for each NE type that adopted SVM and CRF approaches. The ACE datasets were used in the process of evaluation. Based on the results, it was unclear whether or not CRF outperformed SVM or the other way around in ANER. Every NE type is sensitive to various features and every feature plays a key role in the recognition of NE in different levels. The best overall performance of the system in light of F-measure was as follows; 83.50% (ACE 2003), 76.7% (ACE 2004), and lastly 81.31% (ACE 2005).

The authors followed-up their study with a later one Benajiba et al. (2009), where they confirmed the importance of considering language independent and language specific features in Arabic NER. They examined the effect of SVM, ME and

CRF models and the F measure results were 83.34% (ACE 2003), 77.61% (ACE 2004) and 82.02% (ACE 2005).

Furthermore, in Koulali and Meziane (2012) study, the authors brought forward an ANER where they used a combined pattern extractor and SVM classifier, with patterns from POS identified text. The system was able to cater for NEs types using CoNLL conference and using a set of dependent as well as independent features. Nine percent (90%) of the system was trained on ANERCorp data and examined on the remainder. The system was examined using various combinations of features, with the best F-measure result obtained being 83.20%.

Artificial Neural Networks (ANNs)

ANN, Artificial Neural Networks (ANNs) represent an important artificial intelligence technology that is deemed to be a common approach used to machine learning and they have the capability of learning and being trained (Omar et al. 2017).

Studies that relate to ANNs include Mohammed and Omar (2012), who proposed an Arabic language model to extract NER with the use of neural network. They employed ANERcorp and other web resources, to obtain 4 types of NEs, which were person, location, organization and others). The authors conducted a comparison between Decision Tree and Neural Network with the use of the same data. The NNT achieved 92% compared to DT with 87% of precision measurement.

Bayesian Belief Network (BBN)

Another study which has been conducted by Alanazi (2017) using the hybrid method, extracts disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic text in the medical domain. The results of the developed system show that BBN performance is promising with 71.05% overall F-measure. The highest F-measure score was achieved in recognizing disease names with 98.10% while the lowest was in recognizing symptoms with 41.66%.

ii. Semi-Supervised Learning (SSL)

This approach is also referred to as bootstrapping that needs a set of seeds for the initiation of the process of learning. It is an approach with weak supervision and a set of preliminary learning tasks utilized for system training (He & Spangler 2016).

Moreover, an Arabic NER system was developed by Althobaiti et al. (2015) that combined the SS approach and distance learning by training the SS NER classifier on distance learning method. The system aimed to extract person, location and organization NEs in MSA and to upgrade for the extraction of other NEs types. The datasets were obtained from online NEWS, BBC NEWS and ANERcorp. The summary of the studies in literature dedicated to ML-based system for Arabic language is presented in Table 2.7.

Table 2.7 Summary of literature review for ML-base system

Author	Linguistic type	Entity type	Method	Domain	F-measure
Benajiba and Rosso (2007)	ANERcorp	Person, Location, Organization, Miscellaneous	CRF	Political, economic/MSA	65.91
Benajiba and Rosso (2008)	ANERcorp	Person, Location, Organization, Miscellaneous	CRF	Political, economic/MSA	79.21
Benajiba et al. (2008a)	ACE Corpora and ANERcorp.	Person, Location, Organization, Miscellaneous	SVM	Political, economic/MSA	80
Benajiba et al. (2008b)	ACE Corpora and ANERcorp.	Person, Location, Organization, Miscellaneous	SVM, CRF	Political, economic/MSA	80.5
Benajiba et al. (2009a, 2009b)	ACE Corpora and ANERcorp.	Person, Location, Organization and Miscellaneous	SVM, ME, CRF	Political, economic / MSA	80.99
Abdul-Hamid and Darwish, (2010)	ACE 2005, ANERcorp.	Person, Location and Organization	CRF	Political/ MSA	81

To be Continued ...

... Continuation

AbdelRahman et al (2010)	ANERcorp	Person, Location, Organization, Job, Device, Car, Cell Phone, Currency, Date and Time.	CRF, bootstrapping	Political, economic/MSA	81.6
Koulali et al. (2012)	ANERCorp	Person, Location, Organization	SVM	Political/ MSA	83.20
Minaei et al (2012)	NoorCorp	person	CRF	Religious/ CA	89.86
Mohammed and Omar (2012)	ANERCorp, web resources	Person, Location, Organization, Miscellaneous	ANN	Political/ MSA	92
alia.morsi, rafea	ANERcorp	Person, Location, Organization, Miscellaneous	CRF	Political/ MSA	68.05
Zirikly&Diab,2014	Egyptian annotated corpus	Persons, names, locations	SS, CRF	Dielectric Arabic	70.2
Al-Shoukry et al.2015	Online resources	persons, locations, organizations, crime types, dates, times	DTC, feature extraction	Criminal/MSA	81.35
Ayah,&Diab, 2015	Microblogs and Dialectal weblogs	NEs in Dialectal Arabic	CRF	Social media	72.68
M. Althobaiti, 2015	NEWS + BBCNEW, ANERcorp	Persons, location, organization	SS, distant learning	MSA	73.10
Dahan et al. 2015	online newspapers	Person, location and organization	HMM	MSA	74.66
(Alsayadi & ElKorany 2016)	Arabic WordNet ontology (ANW)	Person, location and organization	CRF	MSA	87.86%
(Alanazi 2017)	King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE)	disease names, symptoms, treatment methods, and diagnosis methods	Bayesian Belief Network (BBN)	MSA / Medical	71.05%

Table 2.7 indicates that ANER with ML systems have been increasingly receiving attention from researchers, where different types of ML models have been utilized (e.g., CRF, SVM, ME and HMM). Majority of them were based on the CRF model, with works focused on supervised ML methods using only a few semi-supervised methods, and none unsupervised methods at all for Arabic language. As for other languages, the common features adapted to ANER ML systems with modifications also arise from specific Arabic text characteristics. Such features are based on word-level features, list lookup, word context and linguistic features. Majority of the ML systems were employed on MSA Arabic, with only a few on Classical or Dialectal Arabic, and they also depend on one ML model. As such, further studies should examine the integration of models to obtain optimum performance outcomes.

This need for further studies is compounded by the fact that Arabic language is distinct from other languages as it is characterized by complex morphology and grammar, and majority of the studies that proposed ANER ML systems employed common features that were applied before. This necessitates the development of new models and features that are aligned with the Arabic language's nature to improve the overall performance and capability of ANER ML systems. Unsupervised approaches, like Latent Dirichlet Allocation (LDA), have been used in English NER (Bhattacharya & Getoor 2006; Xu et al. 2009). The LDA has been described as a probabilistic generative model of the text documents for semantic representation by past studies (Blei et al. 2003; Griffiths et al. 2007; Andrews & Vigliocco 2010; Saif et al. 2016). It is based on the premise that each document is a combination of topics. It depends on a set of Dirichlet priors that identify the way document topic combinations may be produced based on latent (random) variables. This may be used in developing Arabic NER to facilitate knowledge acquisition in supervised approaches.

In the past ten years, ANER studies has showed increasing frequency with many of the authors developing ML systems for ANER through the use of established ML models, like CRF, SVM, ME and HMM (mostly CRF). Added to this, works largely focused on supervised ANER ML, while semi-supervised types caught lesser attention, and the unsupervised one caught no attention at all. Most ANER ML system studies also concentrated on the MSA domain, with little attention focused on Classical Arabic

or Colloquial Arabic. They also focused on few NEs types and domains, with other domains like sports, religion, drugs, getting little to no attention.

2.6.3 Hybrid Approach

This approach integrates handcrafted rules (grammar) system with learning-based system for Arabic NER to maximize the complete performance of the system Petasis et al. (2001).

Abdallah et al. (2012) developed and proposed the hybrid NER system for Arabic language to extract names of locations, persons and organizations. They used the licensed linguistics resources of Automatic Contact Extraction (ACE) corpora and Arabic Treebank (ATB) Part 1 v 2.0 dataset, and the free linguistic resource of ANERcorp. Moreover, the rule-based component is a re-implementation of the NERA system employing the GATE tool, whereas the ML-based component employs the decision trees to develop the NE classifier (Shaalan & Raza 2008). The performance indicators using ANERcorp showed the following F-measures for person, location and organization NEs respectively; 92.8%, 87.39% and 86.12%.

The hybrid approach was also used by Oudah and Shaalan (2012) and to contribute to hybrid NER for Arabic language. The system notably managed to extract 11 categories of NEs (names of person, location, ISBN, date, organization, measurement, percent, price, file name, time and phone number). Experiments were carried out through three different ML classifiers to evaluate the overall hybrid system performance. The empirical finding showed that the approach outperformed rule-based as well as ML-based approaches, with the F-measure for person, location and organization NEs using ANERcorp found to be 94.4%, 90.1% and 88.2% respectively.

Meanwhile, in Boujelben et al. (2014) study, the authors also brought forward a hybrid approach that combined ML and rule-based methods to determine the relationships between Arabic words in the text, using manual patterns for complex examples. The first phase entailed the generation of training data set via pre-processing and rule-mining. The third phase used linguistic models that established the handcrafted

types. The types of NEs, which were identified were location, organization and person in MSA text. An F-measure of 75.22% was found on a dataset using ANERcorp.

Moreover, a hybrid approach using CRF method with lexical, contextual features and included up to two adjacent words and the first and the final three characters of one token was proposed by Alotaibi and Lee (2014). They utilized morphological features of person, gender as well as syntactical features of part of speech. They also utilized dependency-based features that specifies the head and dependent words/tokens in the context, and their corpus was WikiFane. The optimal F-measure obtained was 69.68%.

Along a similar line of study, Meselhi et al. (2014) came up with a hybrid Arabic NER approach, combining SVM ML and rule-based system to identify eight NEs types (person, location, organization, date, time, price, percent and measurement. The rule-based approach made use of some grammar rules, whereas the ML approach dependence on extracted features from an annotated text. The findings showed F-measures of 0.983, 0.974, 1.00, 0.987 and 1.00 for NEs date, time, price, measurement and percent respectively.

Moreover, A hybrid approach using integration between semi-supervised and distant learning methods was proposed by Althobaiti et al. (2015). The Independent Bayesian Classifier Combination (IBCC) was used as classifier. The use IBCC to improve the performance. The performance obtained using ANERcorp for F-measure was 70.60% for person, 76.05 % for location, and 72.26 % for organization NEs. The overall F-measure is 73.10. The summary of the studies in literature dedicated to hybrid approach for Arabic language is presented in Table 2.8.

Table 2.8 Summary of literature review for hybrid approach

Author	Dataset	Entity type	Method	Domain	F-measure (%)
(Abdallah et al. 2012)	ACE 2003, ANERcorp	Person, Location, Organization	Rule-based with CRF	MSA/ Politic	88.77
(Oudah & Shaalan 2012)	ACE, Treebank, ANERcorp	Person, Location, Organization, Date, Time, Price, Percent, Phone Number, Measurement, ISBN, File Name	Rule-based with SVM, DT, (J48), Logistic Regression	MSA/ Politic, economic	90.9
(Boujelben et al. 2014)	ANERCorp	Person, location, organization	Linguistic models with supervised ML	MSA	75.22
(Alotaibi & Lee 2014)	WikiFane	Miscellaneous	CRF, dependency rule-based features	MSA	69.68
(Meselhi et al. 2014)	ACE 2005	person, location, organization, date, time, price, precent and measurement	Rule-based system with SVM	MSA	99
(Althobaiti et al. 2015)	(NEWS + ANERCorp)	Person, location, organization	semi-supervised and distant learning techniques	MSA	73.10

2.7 LITERATURE RESEARCH GAPS

The development in Arabic language processing research are still in its infancy in comparison to other languages, like English. These may be attributed to the challenges inherent to the Arabic language compounded by the lack of annotated corpora and resources. Added to this, prior studies presented many factors when addressing this issue with the entity factor revealed as one of the determinants of NER and several entities have been mentioned and used in studies.

More specifically, in NER studies, person NEs recognition is a top research category with many studies dedicated to it. In Arabic language, one of the significant studies that focused on person NEs was Shaalan and Raza (2007), who utilized rule-based approach with linguistic grammar-based techniques and whitelist dictionaries to accurately recognize person NEs in Arabic text. They considered three components and they were name lists (gazetteer) regular expressions (grammar) forming the lexicon, and

filtration mechanism via the establishment of grammatical rules that assisted in sifting through invalid NEs.

Similarly, Elsebai et al. (2009) used a rule-based approach using pattern matching with morphological analyzer, and Aboaoga and Ab Aziz (2013) used the same approach on person NEs in Arabic. On the other hand, Bidhend et al. (2012) adopted a different approach for sole person NER system. They made use of ML based approaches with CRF method, while Al-Shalabi et al. (2009) proposed an Arabic NER algorithm to retrieve Arabic proper nouns through the use of lexical triggers. They took into account regional patterns (e.g., name connectors) and the algorithm identified NE types of people, major cities, locations, countries, organizations, political parties and terrorist groups. The algorithm is directed by heuristic rules for pre-processing of the input to clean and remove affixes from the data. Internal evidence triggers of person name connectors were then used for NEs recognition.

More importantly, research on NER systems has been extended to determine and recognize various types of NEs aside from person NEs. Majority of the studies brought forward systems that are capable of extracting three types of NEs, which are location, person and organization. For instance, Zaghouani (2012) used the rule-based on RENAR system divided into three levels (morphological pre-processing, known name identification and application of local grammar) to determine unknown NEs. In another study, Abdul-Hamid and Darwish (2010) focused on all the three mentioned NEs by using a simplified features set for ANER. Their system had its basis on CRF to recognize the NEs types and only consider surface features (leading and trailing character n-gram, word position, word length, word unigram, probability, the preceding and succeeding words n-gram and character n-gram probability) excluding any other feature type.

Moreover, an Arabic NER with a combined pattern extractor and SVM classifier that learns POS tagged text patterns was developed by Koulali and Meziane (2012). They covered the three NE types and used a group of dependent and independent language features. Along a similar line of study, Abdallah et al. (2012) used an NER hybrid approach to recognize the above three NE types for Arabic.

Contrastingly, in some studies on NER, systems were developed with the capability of recognizing and extracting NEs aside from the above three types. This is attributed to the pioneering work by Maloney and Niv (1998), who proposed a rule-based approach that was capable of identifying five NEs (person, location, organization, number and time). Mesfar (2007) further extended the types of NEs by adding currency and temporal expressions, and Shaalan and Raza (2007) proposed a rule-based NER system that was capable of recognizing 10 types of NEs including, date, time, ISBN, price, measurement, phone numbers and file names. Also, Traboulsi (2009) rule-based system identified time, date and address expression that is of great benefit for identifying correspondent addresses.

In some other studies, NER systems that are able to recognize miscellaneous NE types were proposed using machine-learning approach Benajiba et al. (2007). Similarly, AbdelRahman et al. (2010) came up with an integrated machine-learning method to identify Arabic NEs, which included job, device, care, cell phone, currency, date and time, whereas Oudah and Shaalan (2012) proposed a hybrid NER system to recognize 11 types of Arabic NEs with the addition of more NEs types like measurement and price.

Another determinant of NER is domain factor and different NER systems have been proposed for certain domains of Arabic texts like MSA or CA. Additionally, some NERs concentrated on a single sub-class of NEs type like politics, economics, sports and religion. Political NEs in MSA text based on rule-based approach were developed by based (Maloney & Niv 1998; Mesfar 2007; Al-Shalabi et al. 2009), and machine-learning approach was revealed by (Abdul-Hamid & Darwish 2010; Koulali & Meziane 2012; Mohammed & Omar 2012; Althobaiti et al. 2015; Dahan et al. 2015; Alsayadi & ElKorany 2016). Moving on to the hybrid approach for NEs recognition in the political domain, combined rule-based with classification (Abdallah et al. 2012). However, NEs in religious domain in light of CA text were largely untouched by studies, with a few daring to delve into the subject (Bidhend et al. 2012).

Meanwhile, in other studies in literature, NER systems that recognize NEs in the domain of politics and economics of MSA texts using rule-based approach were

reported by (Shaalan & Raza 2007; Elsebai et al. 2009; Shaalan & Raza 2009; Traboulsi 2009; Zaghouani 2012) designed NER systems for the domain of politics and economics. Meanwhile, (Benajiba et al. 2007; AbdelRahman et al. 2010) brought forward machine-learning approaches that are capable of identifying NEs in the above two domains, while Oudah and Shaalan (2012) did the same but used a hybrid approach. That means hybrid method have been tried in other domain except the religion. Thus, there is a need to enhance the result in religion domain by developing hybrid approach.

Literature also reports NER systems that have the capability of recognizing NEs in over two domains; for instance, aside from political and economic domain, Aboaoga and Ab Aziz (2013) introduced rule-based approach to recognize NEs in sports domain in MSA texts. Literature also reports NER systems that have the capability of recognizing NEs for criminal domain (Asharef et al. 2012; Al-Shoukry & Omar 2015) introduced ML approach to recognize NEs.

More recently, new study in literature, NER system that recognize NEs in the medical domain of MSA texts using ML approach were reported by Alanazi (2017).

Still another NER determinant is the linguistic resource factor. In the case of Arabic language, its processing is quite challenging because of the scarcity of resources and limited Arabic text data sets that are available and capable of offering functionalities and testing platforms for Arabic NEs systems and applications. This was headed by the pioneering work of Maloney and Niv (1998) referred to as TAGARAB that combines a pattern-matching engine and supporting data have the morphological analysis element. Following their study, several Arabic datasets were made available for use like NooJ linguistic environment by Mesfar (2007), ACE and Treebank Arabic Dataset used by Shaalan and Raza (2007). Moreover, ANERcorp is the most extensively utilized Arabic resource in literature, which was employed by (Benajiba et al. 2008; Elsebai et al. 2009; Abdallah et al. 2012; Mohammed & Omar 2012; Oudah & Shaalan 2012; Zaghouani 2012; Boujelben et al. 2014; Althobaiti et al. 2015) in rule-based and machine learning for NER systems, ANERcorp was also extensively utilized – it was also utilized in the testing platform.

Along with the above-mentioned resource, Arabic Corpus is also used to test the rule-based NER system in Traboulsi (2009) study. In relation to this, more than one dataset type was used by some researchers to test their system; for instance, Benajiba et al. (2008), Benajiba et al. (2009) and Abdul-Hamid and Darwish (2010) used ACE Corpora and ANERcorp in their machine learning-based NER system. Other studies like Abdallah et al. (2012) used ACE 2003 and ANERcorp in their hybrid system whereas in their pipelined NER hybrid system, Oudah and Shaalan (2012) used both ACE 2003 and ANERcorp coupled with Treebank.

Generally, the progress in NLP processing research is still in its infancy in comparison with other languages and this may be attributed to the challenges in Arabic language and the scarcity of annotated corpora and resources. It is however notable that the works dedicated to rule-based NER systems that is a significant research topic in Arabic NLP have shown considerable developments to date but more progress is needed and efforts of further research are called for as majority of the proposed rule-based systems focused on MSA, with CA receiving little to no attention. CA is basically related to the domain of religion and ancient Arabic literature and poetry which is a good avenue for research in Arabic NER. Opportunities for research can be taken up in developing new handcrafted rules that can use Arabic grammar, particularly because rule-based NER for MSA texts are still confined to few types of NEs and fewer domains. Hence, further research should look into developing rule-based NER systems as called for by prior studies (Oudah & Shaalan 2012) that introduce NER novel domains like criminal records, drugs, sports, religion, among others.

In other studies, ML-systems have been developed for ANER using well-validated ML models (CRF, SVM, ME and HMM), with most gravitating towards CRF. Also, majority of the works of this caliber concentrated largely on supervised ANER ML studies, with only few on semi-supervised types, and none on unsupervised ones. Additionally, ANER ML systems studies are largely confined to MSA domain, with only a few on Classic and Colloquial Arabic. To compound the need for further studies, ML NER for MSA texts address a few types of NEs and domains, with other domains remaining under-investigated (e.g., criminal records, sports, religion, drugs). Owing to the scarcity of Arabic language corpora, a few focused on the Arabic language and this

underlines the need for the development of new Arabic corpora to function as a platform for new studies in different domains.

2.8 SUMMARY

The influence of NER has caught the attention of several researchers in the area of NLP. NER is basically an important task in many Arabic NLP applications and it is useful to use in many tasks including Information Extraction (IE), Question Answering (QA), Information Retrieval (IR), as well as Machine Translation (ML). NER employing applications is a significant pre-processing phase that is conducted to improve the overall performance. In relation to this, the sixth Message Understanding Conference (MUC-6) was the first to introduce the NER task.

This chapter provided a discussion of relevant studies in literature concerning the fundamental aspects of NLP and NER and the challenges of the Arabic language. The chapter also contains discussions on past studies concerning rule-based approach, machine learning approach and hybrid approach. It concludes with the challenges and gaps in prior literature with the aim of development the research problem.

CHAPTER III

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter presents a full description of the research methodology for this study. The chapter is based on the findings of a review of the relevant previous works on the ANER. The review identified the most relevant and suitable methods that are available to achieve the research objectives. The aim of this chapter is to provide the appropriate pathway to achieve the objectives of this study. Therefore, first a description of the design-based research (DBR) methodology adopted in this study is provided in Section 3.2. A detailed discussion of the phases of the research methodology is presented in Section 3.3. A description of the tools used to prepare the data and to run those data on the rule-Based and machine learning algorithms are presented in Section 3.4. A Baseline analysis for GATA and Language Computer in Section 3.5. Finally, a summary of this chapter is presented in Section 3.6.

3.2 DESIGN-BASED RESEARCH (DBR)

Design-based research is also commonly known as design research, development research, and design experiments (Barr & Wells 1990; Barab & Squire 2004; Parker 2011). The DBR methodology was presented by a small group of scholars in the DBR collective as a way to develop algorithms and systems (Kelly 2003; Barab & Squire 2004; von Alan et al. 2004). The methodology is an iterative research process that includes analysis, design, development, and implementation in real-time settings (Wang & Hannafin 2005; Parker 2011). Design-based research is a systematic but flexible methodology that was proposed to help bridge the gap between research and practical implementations (Aken 2004; Oates 2006). The key aspects of the method include the

handling of complicated issues, integrating design principles with new technologies to develop practical solutions to the problems, and conducting effectiveness evaluations to improve the proposed solutions and to identify new design principles (Reeves 2006; Parker 2011). Figure 3.1 illustrates the four main phases of the DBR methodology.

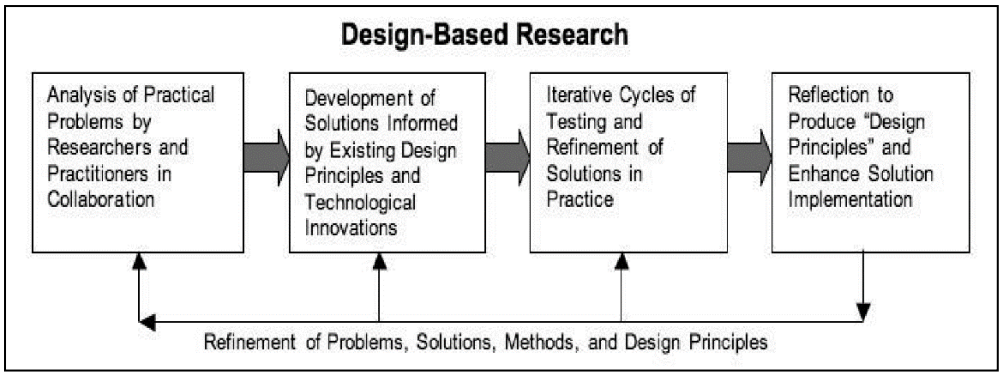

Figure 3.1 Four phases of design-based research methodology

In line with Figure 3.1, Parker (2011) states that the DBR methodology consists of four iterative phases, which are described in the following sections.

Phase 1 - Analysis of Practical Problems by Researchers and Practitioners in Collaboration

In this phase, the problems and the literature review are addressed. During this phase, the problems are clearly identified and investigations are conducted into the work that has already been carried out in the same or related fields. By the end of this phase, it should be possible to determine the preliminary research questions and objectives to guide the research (Herrington et al. 2006).

Phase 2 - Development of Solutions Informed by Existing Design Principles and Technological Innovations

This phase focuses on designing and developing solutions to the problem. During this phase, a more targeted literature review is conducted. Relevant theories as well as design principles for ANER approaches are explored and developed in depth in order to propose a new approach to solve the research problem (Herrington et al. 2006).

Phase 3 - Iterative Cycles of Testing and Refinement of Solutions in Practice

In this phase, a potential solution to the problem is implemented and tested internally in an iterative manner to assess its effectiveness and correctness (Reeves 2006).

Phase 4 - Reflection to Produce 'Design Principles' and Enhance the Implementation of the Solution

In this phase, the proposed solution that is implemented needs to be comprehensive, and this is confirmed by evaluating it against existing related works (Herrington et al. 2007).

3.3 ADOPTED RESEARCH METHODOLOGY

The stages, features, and attributes of the DBR methodology are customized and adopted as the research methodology for this study. Figure 3.1 presents the four interrelated phases of this study's research methodology. These phases describe the research objectives/activities undertaken.

Although the activities in the methodology seem to be a linear process, the iterative process is still applied as the literature review, theoretical knowledge, and ANER approaches.

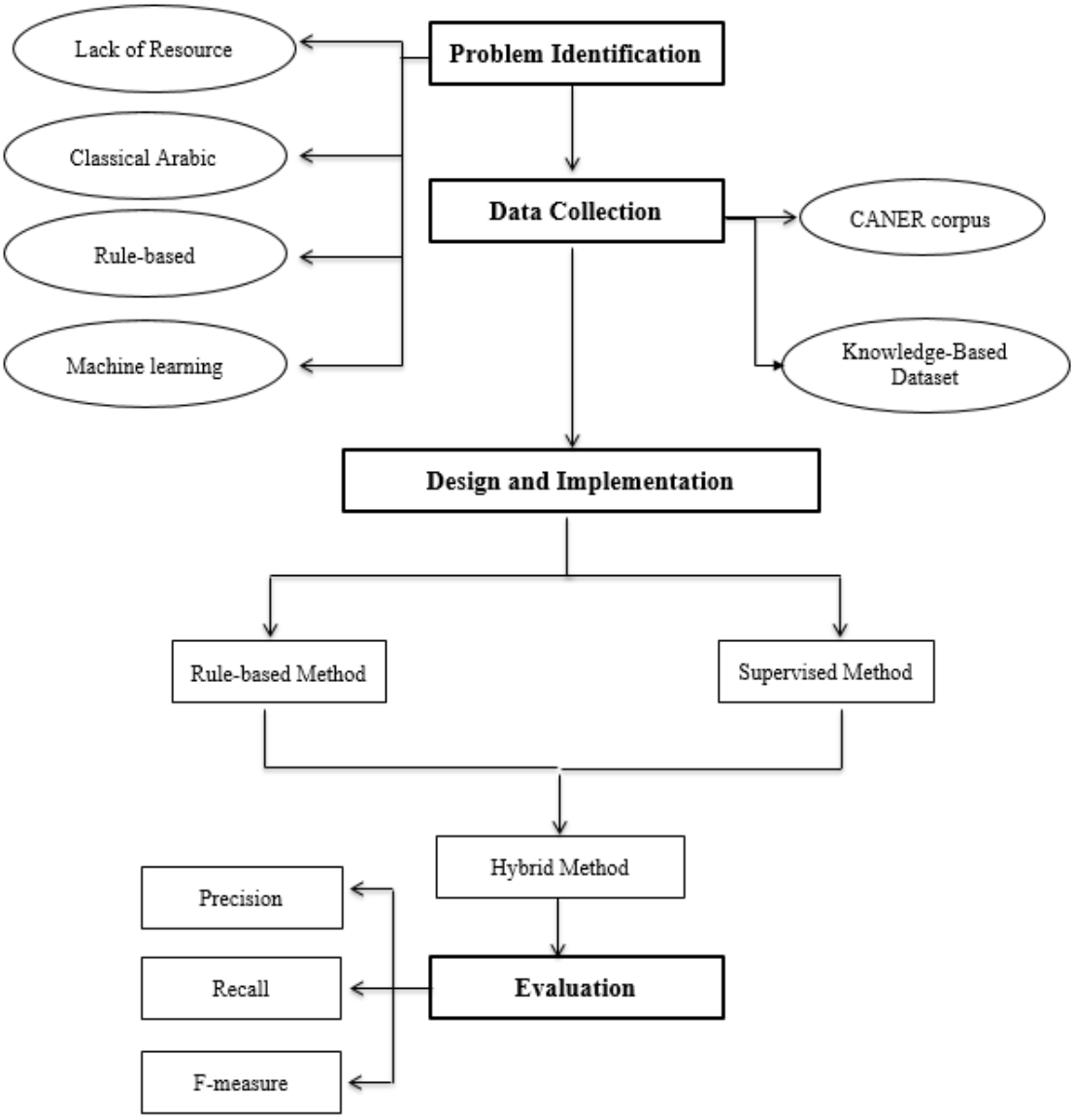

Figure 3.2 Phases of research methodology

Figure 3.2 shows the process of the methodology phases started with problem identification which includes, lack of resource, Rule-based approach, Machine learning method, and Classical Arabic. The details of each phase are explained in the following sections.

3.3.1 PHASE 1 – Problem Identification

The reviewing process is considered as an elementary and essential phase in this research where the research framework is identified based on the observations made on the previous related works that have been conducted for ANER. As shown in Figure 3.1, three directions of research are highlighted based on the analysis of important features that include challenges, problems, and research objectives with their intended outcomes.

Arabic NER task faces many serious problems that spread over all the levels of processing from the first stage until the final stage of recognition. These problems are determined in each stage of NER in Arabic. Literature review shows that a large number of works have been investigated and developed for Arabic NER using three main methods. It also revealed that NER in Arabic context is not a trivial task due to the complexity of the Arabic language (Abdallah et al. 2012; Shaalan et al. 2012; Salah & binti Zakaria 2017).

The rule-based methods (Mesfar 2007; Elsebai et al. 2009; Halpern 2009; Traboulsi 2009; Shaalan 2010; Zaghouani 2012; Aboaoga & Ab Aziz 2013; Elsayed & Elghazaly 2015) use the morphological, grammatical, syntactic, and semantic information to extract and identify the NEs in Arabic texts. Most of these methods depend on the gazetteers or key-words which occur before or after the named entities in Arabic texts. However, they are unable to deal with the complex structures of Arabic texts, as well as Arabic NEs. Furthermore, the rule-based methods cannot identify some Arabic NEs which occur without identifier (trigger words) and are not listed so far in the gazetteers such as the old NEs that doesn't used any more.

The supervised methods depend on the tagged corpus that is used for training the classification techniques such as Naïve Bayes and support vector machine. Unfortunately, when NER algorithms are trained on such texts, they tend to perform poorly on shorter. The problem stems from the very limited amount of CA gold standard datasets currently available (Bontcheva et al. 2017). Although a few corpora Benajiba et al. (2007) have been collected for Arabic NER, these corpora are considered a general

domain sources, as well as MSA. The CA is important due to its association with Islamic domain. In the best of the author's knowledge, there is no tagged corpus for NER of classical Arabic. In addition, CA has its special named entities such as Allah, prophet, religion, sect, paradise and hell, which require a specific solution for identifying them in the Islamic domain.

Generally, one of the main problems facing the supervised methods is that they are unable to deal with low-features of NEs in the training step. In fact, most of the words (features) that appear before/after NEs in the corpus may have low frequencies. This means that the supervised methods often exclude many important NEs with lower frequencies. This problem will have a significant impact in the Arabic language because Arabic is an agglutinative language with very rich morphological variations. Several features are examined in the existing supervised methods such as stop words, stemming, and part-of-speech. The main problem is how to select the informative features by using language reduction methods in order to improve the performance of Arabic NER. This problem arises more from the misleading words that should be ignored from the beginning of pre-processing stage, where the high dimensionality of text should be reduced without degrading the performance of supervised techniques and without loss the relevant information during reduction process.

In recent years, there are a few number of works which have been introduced for Arabic NER using the hybrid method approaches (AbdelRahman et al. 2010; Abdallah et al. 2012; Oudah & Shaalan 2012; Meselhi et al. 2014; Meselhi et al. 2014; Shaalan & Oudah 2014; Alanazi 2017) to overcome the limitations of the rule-based and supervised methods. In the hybrid methods, the serious challenge in identifying NEs is to construct logically sound models for describing the different types of NEs, because the NEs have a huge variety of linguistic features and include a wide range of phenomena, from simplex words to sentence patterns. Usually, the incorporate of several techniques can achieve better performance than both techniques in one framework. However, the main challenge is how to incorporate the advantages of rule-based and supervised method together and overcome their weakness to create a hybrid method for Arabic NER.