Journal of Theoretical and Applied Information Technology
31st December 2018. Vol.96. No 24

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

BUILDING THE CLASSICAL ARABIC NAMED ENTITY RECOGNITION CORPUS (CANERCORPUS)

[1]RAMZI ESMAIL SALAH, [2]LAILATUL QADRI BINTI ZAKARIA

[1]SamaaSoft Company, Department of Research Affairs and Innovation, Saudi Arabia

[2]Universiti Kebangsaan Malaysia, Department of Computer Science, Malaysia

E-mail: [1]ramzi@samaasoft.com, [2]lailatul.qadri@ukm.edu.my

ABSTRACT

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 natural language processing applications that require human annotation. In the Named Entity Recognition (NER) task, several researches have been introduced to address the complexity of Arabic in terms of morphological and syntactical variations. However, there are a small number of studies dealing with Classical Arabic (CA) that 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 of great value if extracted. Therefore, in this paper, we introduce Classical Arabic Named Entity Recognition corpus as a new corpus of tagged data that can be useful for handling the issues in recognition of Arabic named entities. It is freely available and manual annotation by human experts, containing more than 7,000 Hadiths. Based on Islamic topics, we classify named entities into 20 types which include the specific-domain entities that have not been handled before such as Allah, Prophet, Paradise, Hell, and Religion. The differences between the standard and classical Arabic are described in details during this work. Moreover, the comprehensive statistical analysis is introduced to measure the factors that play important role in manual human annotation.

Keywords: *Arabic corpus, Corpus generation, Named entity, Classical Arabic*

1. INTRODUCTION

A named entity is a term or word that clearly identifies an object from a set of other objects with similar traits. The term "Named Entity" was introduced at the Message Understanding Conference (MUC-6) in 1995 [1]. In the expression named entity, the word named limits the scope of entities that have one or many rigid designators that stands for a referent. Usually, rigid designators include proper names, but it depends on the domain of interest that may refer the reference word for object in domain as named entities. For example, in molecular biology and bio-informatics, entities of interest are genes and gene products. The dominant classification of NE in previous studies [2-4] includes three classes, such as Person, Location and Organization names. However, there are several specific domains for named entities that occur in the certain domain such as diseases and medicine in biomedical domains [5].

Named entity recognition (NER) is an important task in natural language processing applications for detecting named entities (NEs) in natural language documents [6]. Arabic NER systems are facing some challenges that are associated with Arabic language such as, capitalization issue [6], complex morphology [7] and lack of resource[7, 8]. Several studies have been introduced for Arabic named entities using three main approaches, rule-based [9, 10], machine learning [11, 12] and hybrid approaches [3, 13]. The main issue that faces the supervised Arabic named entity recognition is knowledge elicitation bottleneck and the lack of resources for underdeveloped languages that require extensive effort from the linguists. In order to conduct an experimental analysis or build efficient systems, textual data would always be critical. The data is required to have some meta-data to denote the phenomena being investigated in a given study. A set of textual data is called a corpus and adding annotation produces an annotated corpus. In [14], manual examination of corpus, automatic analysis of corpus, reusability of annotations, and multi-functionality were identified among the significant factors of corpus creation. Creating a corpus that normally requires a combination of hard work from a human linguist

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

who needs to label each word with its corresponding class. On the other hand, evaluating the Named Entity Recognition (NER) method in Classical Arabic (CA) domain requires or needs a gold standard data set that contains different types of named entities which are related to CA. To the researcher's best of our knowledge, there is no gold standard (available evaluation data sets) for the NER in CA in the literature review of previous research [15]. Therefore, in this study, an Arabic data set was created for evaluating the proposed method. The knowledge source that is used as the background information for creating the data set is introduced as well.

This study will be dedicated to Classical Arabic Named Entity Recognition Corpus (CANERCorpus) as a new tagged data that can be useful for handling the issues in recognition of Arabic named entities. It is a freely available (https://github.com/RamziSalah/Classical-Arabic-Named-Entity-Recognition-Corpus) and manual annotation by human experts and contains more than 7,000 Hadiths. Based on Islamic topics, named entities are classified into 20 types which include the specific-domain entities that have not been handled before such as Allah, Prophet, Paradise, Hell, and Religion. The differences between the standard and classical Arabic are described in detail during this work. Moreover, the comprehensive statistical analysis is introduced to measure the factors that play an important role in manual human annotation. The paper presents the methodology and results of the CANER corpus.

2. RELATED WORK

The past decade has witnessed construction of the background information resources to overcome several challenges in text mining tasks [7]. For non-English languages with poor knowledge sources such as Arabic [16], these challenges have become more prominent especially for handling the NLP applications that require human annotations [17]. In the NER task, several researches have been introduced to address the complexity of Arabic in terms of morphological and syntactical variations [18]. 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. [19], 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 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. [19] 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 encyclopaedia 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 [20, 21], word sense disambiguation [22, 23], building or enriching lexical sources [4, 24] and NER.

Another works by Mohit, et al. [25], 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) and miscellaneous.

Attia, et al. [26] 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 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 [27] described a supervised machine

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

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. [28], 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.

King Saud University Corpus of Classical Arabic (KSUCCA) [29]. KSUCCA is a pioneering 50 million words corpus of Classical Arabic with various genres and sub genres that can be used in various types of Linguistic and Computational Linguistic research, but it's not focusing on the NE.

Recently, the knowledge-based approach by and [4, 30] 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, as well as some organisations and geographical locations names [15]. There is also some automatic creation of Arabic named entity annotated corpus with small size in modern Arabic.

Most of the previous work on Classical Arabic was not Named Entity corpus-based, even if they are, the classification of names is automatically annotation, so the error is more frequent, or they do not focus on the NE but focus on the Classical Arabic language. In this study, named entity corpus in classical Arabic that focuses on Islamic domain is created to satisfy the need for a new corpus, we focused on the Classical Arabic, and classify all the corpus names or others manually with evaluation and verification of accuracy of information, 20 type of names is considered more than 6 of them are new.

3. METHODOLOGY FOR BUILDING THE TAGGED CORPUS OF ARABIC NER

Figure 1 shows graphically the main phases of the methodology for building the tagged corpus of Arabic NER. The first phase focused on compiling and preparing the knowledge source as background information for the NER corpus in the Islamic domain. After compiling the knowledge source, the NE types in the Islamic domain are identified to form the theoretical observations for the NER corpus. The third phase aimed to prepare the knowledge source in order to use it in the next phase. After that, each document was segmented into its sentences. Then, the sentences were introduced to the human experts to label the NE based on the NE types list. Finally, the evaluation phase was performed to evaluate the human judgments in this work, and the results obtained in this phase were also analyzed.

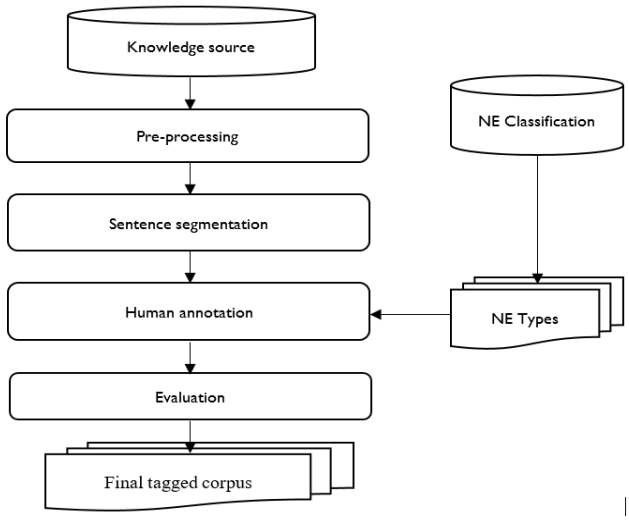

Figure 1: Methodology of building CANER corpus

3.1 Knowledge source

The knowledge source of the data set was collected from the Islamic contexts because the focus of the current research is on the NER in CA. It is based on the most authentic and widely used six collections which are known as Ṣaḥīḥ al-Bukhārī [31].

Prophetic traditions (Hadith) have been compiled by different Muslim scholars. Among them are the six collections which are known as Saha Satta (the authentic six). Ṣaḥīḥ al-Bukhārī (Arabic: صحيح البخاري) is one of the main authentic and trusted

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

Hadith collections after the holy Quran. It contains conversations and sayings of the Prophet Mohammad (peace be upon him) which are narrated by Prophet's companions and other authentic narrators through single or multiple narration lines. Ṣaḥīḥ al-Bukhārī was collected by Imam Muhammad bin Ismail Al-Bukhari. It is considered as the first collection of prophetic traditions Bukhari [32]. This collection is the most prominent book that compiles prophet Muhammad's Hadith which is regarded by the Sunni Community as the second important Islamic resource after the holy Quran [31]. This book covers almost all aspects of Prophet Muhammad's life. It also provides Muslims with proper guidance to Islam, including the method of performing prayers and other acts of worship. In his book, Bukhari classified prophetic narration concepts which consist of three hierarchical levels. Sahîh Al-Bukhari contains more than 7,000 Hadith. Imam Bukhari completed this book in the year 232 of the Prophet's migration. TABLE I presents the statistical information about Sahîh Al-Bukhari.

Table 1: Statistical dataset

Word	Value	Word	Value
Size	8249KB	Lines	20546
Items of Hadith	7124	Words	258264
Pages	1345	Named Entity	72,108
Paragraphs	12354	Others	186,133

All writings in Sahîh Al-Bukhari are in CA and they focus on the Islamic domain. The number of pages of this Islamic corpus is 3172. We collected this corpus from different places in Sahîh Al-Bukhari, and the overall number of words in our corpus is 258264. We used these words and extracted twenty types of NEs.

3.2 Pre-processing

One of the most important tasks prior to many NLP applications is preparing the data [33]. This is more important especially when the data is collected from websites. In this work, the function of the pre-processing task is deleting all extra spaces or diacritics. In addition, we use normalization to reduce different forms of words into one form because in Arabic, one word may be written in many forms, for example, (“Alef”) (أ,ء,ؤ,ئ,ى) to the normal way. We will process data to be free of the ambiguity principle.

3.3 Sentence segmentation

The preliminary step of annotating corpus is commonly performed on texts before further processing it. This step aims to identify the sentences boundaries and to split the running text into tokens.

3.4 Human annotation

The manual tagging annotation for the compiled data set is accomplished by three human annotators (experts in computational linguistics) annotated the corpus over a period of 1-year, part time. Thus, all the experts have more than five years of experience in the field of linguistics. Each document was segmented into sentences. The types of NE were also presented to provide the annotators with sufficient information to make a decision. For each sentence, the annotators were asked to determine the NEs and label them as one of the categories that will be described later. The ‘O’ category indicates that the words in the sentences are not NEs.

3.4.1 NER Annotation

There have been various schemes for the annotation format of corpora for different NLP tasks. The present study uses two main schemes in the NER annotation, which are inline annotation and standoff annotation.

Column-based format is a simpler form of annotation which places each word on a single line with its corresponding class delimited by a tab or a space. It is sufficient when there is no nested annotation required by the task definition, such as POS tagging. According to Kudo and Matsumoto (2001), two schemes were used in text chunking: “inside/out” (Ramshaw & Marcus 1995; Sang & Veenstra 1999). It is based on annotating a token with its position (P) either inside a named entity or outside it, and attaching that position to its entity class (CCC). The full label would have the form (P-CCC). One generic outside class was used for all (O). To mark entity boundaries, (P) would take (B) at the beginning boundary of an entity in the IOB scheme and (E) then, the ending boundary in the IOE scheme is marked.

- The IOB schemes have the following variants:
  - IOB1: assigns B only if it is followed immediately by another token of the same entity type.
  - IOB2: assigns B whenever it starts a new entity; head entity.
- The IOE schemes have the following variants:

(1) IOE1: assigns E for the end token if it is immediately preceded by another token of same type.

(2) IOE2: assigns E whenever it an entity is ended.

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

"start/end" In this scheme, S was added and to all tags were used inside/outside the schemes to represent single token entities. Then, B and E tags were assigned regardless of the preceding token class. Table 2 presents examples illustrating all the discussed schemes.

Table 1: NER annotation schemes example

Token	IOBI	IOB2	IOE1	IOE2	Start/End
Ramzi	I-PER	B-PER	I-PER	I-PER	B-PER
Salah	I-PER	I-PER	I-PER	I-PER	I-PER
Travelled	O	O	O	O	O
To	O	O	O	O	O
Meet	O	O	O	O	O
Saleh	I-PER	B-PER	I-PER	I-PER	S-PER
In	O	O	O	O	O
Mecca	I-LOC	B-LOC	E-LOC	E-LOC	S-LOC
Saudi Arabia	B-LOC	B-LOC	I-LOC	E-LOC	S

Different schemes were used without a decision as to select the best one [34]. The most widely used is IOB2 but classifiers built on different schemes [35]. IOB2 was adopted for CoNLL evaluations and the data consists of two columns separated by a single space. Each word was put on a separate line and there is an empty line after each sentence. The first item on each line is a word and the second is a named entity class. The tags used in CoNLL were person names (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). In this work, IOB2 was followed.

3.5 Evaluation

In order to evaluate the results that obtained from humans' judgments in this work, the word in each sentence was assigned to a certain category which is marked by at least two annotators as a category. However, when the annotators selected three different categories, we just put such tokens as O category. Reliable data is important for researchers who wish to use manual tagging annotation with categories, whether to support an empirical claim or to develop and test a computational model. In this paper we assume that the data is reliable if annotators agree on the categories assigned to tokens [36, 37]. If different annotators give similar results consistently, then we can deduce that they have gotten a similar understanding of the guidelines, and we can expect them to perform under this understanding. Reliability is therefore a prerequisite for the validation of the annotators. If the annotators are not consistent then either one or more of them are wrong. The simplest measure of agreement between two annotators is percentage of agreement or observed agreement. This is the number of items on which the annotators agree divided by the total number of items. More precisely, and looking ahead to the following discussion, observed agreement is the arithmetic mean of the agreement value arg_i for all items $i \in I$ [38], defined as follows:

$$arg_i = \begin{cases} 1 & \text{if the two annotators assign } i \text{ to the same category} \\ 0 & \text{if the two annotators assign } i \text{ to different categories} \end{cases}$$

Observed agreement over value arg_i for all items $i \in I$ is then:

$$A_o = \left(\frac{1}{r}\right) \sum_{i \in I} arg_i$$

TABLE 3 shows annotations per class. It also demonstrates how the three annotators (A, B, C), in many cases, agree with each other. Moreover, the majority is the actual NE number for every class.

Table 3: Annotations per class

Class	A	B	C	Majority
Allah	8143	7710	7823	7811
Prophet	6108	6326	6420	6502
Pers	39243	39201	39230	39159
Loc	1324	1271	1279	1349
Org	9	9	9	9
Meas	148	146	141	147
Mon	139	139	139	139
Book	182	189	179	183
Date	595	609	599	596
Time	102	112	104	102
Rlig	184	184	184	184
Sect	15	21	17	17
Clan	663	676	691	674
NatOb	668	668	668	670
Crime	200	209	214	212
Para	294	294	294	294
Hell	245	245	245	245
Month	77	77	77	77
Day	31	31	31	31
Num	13707	13707	13707	13707

Table 3 presents the reliability of the annotators' decision which was computed as the average of the pairwise observed inter-annotator agreement A_o [38]. According to TABLE II, some named entities obtained a decision with majority value, reflecting the annotators' agreement such as number, while other named entity yielded different opinions due to the relation with other named entities such as person. The percentage agreement between annotators is shown in TABLE 4.

Table 4: Inter-annotator agreement

Class	A-B	A-C	B-C
Allah	94.6825%	96.07%	98.55%

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

Prophet	96.5539%	95.14%	98.53%
Pers	99.8930%	99.96%	99.92%
Loc	95.9970%	96.60%	99.37%
Org	100%	100%	100%
Meas	98.64%	92.56%	93.83%
Mon	100%	100%	100%
Book	97.06%	89.24%	91.94%
Date	88.62%	90.15%	98.30%
Time	91.07%	98.07%	92.85%
Rlig	100%	100%	100%
Sect	71.42%	88.23%	80.95%
Clan	98.07%	95.94%	97.82%
NatOb	100%	100%	100%
Crime	95.69%	93.45%	97.66%
Para	100%	100%	100%
Hell	100%	100%	100%
Month	100%	100%	100%
Day	100%	100%	100%
Num	100%	100%	100%

4. NAMED ENTITY CLASSIFICATION

The number of NE types has been quite limited [39]. Based on Modern Standard Arabic (MSA), most previous studies in Arabic NE recognition focused on these named entities: personal name, location and organization [15]. Many more kinds of things have proper names or proper classes of expressions [39]. The CA has its privacies on classifying NEs especially in the Islamic domain. Therefore, classification of NEs must take into account privacies of the CA. In this work, as shown in Figure 2, the NE was classified into two main types:

The general type covers persons, locations, organizations, measurement, money, book, date, time, natural object, crime, day, and number where you can find this type in many domains such as politics, economy, sport, and crime, and others.

The second type known as the specific domain is related to CA (Islamic domain), which includes Allah, prophet, religion, sect, paradise and hell. However, the context of the corpus that includes general and specific NEs focuses on the Islamic domain. Therefore, there are many differences in the names, meanings and roles between the Islamic domain and other domains.

These group of Islamic domains has been chosen due to the following: small data set of Hadiths; these data had been analyzed by three experts to identify the famous NE in CA; the experts found the six types that are more general in CA which are Allah, prophet, religion, sect, paradise and hell; those NE are also well known in Islamic Religion as well.

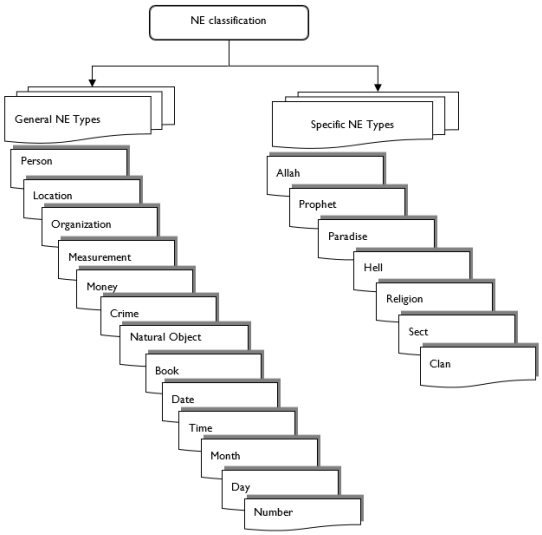

Figure 2: Classification of NE

4.1 NE Types for general domain

The first one is general classification (person, location, organization, measurement, money, book, date, time, natural object, crime, day and number) which can be found in many domains such as politics, economy, sport, and crime, etc.

TABLE 5 presents the types of NE detected for the general domain. As reported in Conference on Computational Natural Language Learning in 2003 [40], a proper corpus for NEs task should include words with their associated NER tags. Thus, we have two labels for each class (B and I) and one O label to denote the "OTHER" class. CANERCorpus contains more than 250,000 tokens tagged according to the IBO2 annotation:

Table 5: Tags definitions for Public Domain

Type	Short	Definition	Example
Person	B-Pers	The beginning of a person name	أحمد/ Ahmed
	I- Pers	The continuation (inside) of a person name.	رمزي/ Ramzi
Location	B-Loc	The beginning of a location name	تهامة/Tahamah
	I-Loc	The inside of a location name	اليمن/ Yemen
Organization	B-Org	The beginning of an organization name	دار/ House
	I-Org	The inside of an organization name	طوق النجاة/ Collar Deliverance
Measurement	B-Meas	The beginning of a measurement name	القدح/ the mug
	I-Meas	The continuation (inside) of a measurement name	

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

Money	B-Mon	The beginning of a money name	دينار/ Dinar
	I-Mon	The continuation (inside) of a money name	
Crime	B-Crime	The beginning of a Crime	السرقة/ Theft
	I-Crime	The continuation (inside) of a Crime	
Natural Object	B-NatOb	The beginning of a Natural Object	حماراً/ Donkey
	I-NatOb	The continuation (inside) of a Natural Object	
Book	B-Book	The beginning of a book name	كتاب/ The book
	I-Book	The continuation (inside) of a book name	الزكاة/ of Zakah
Date	B-Date	The beginning of a date	يوم/ Day
	I-Date	The continuation (inside) of a date	الفيل/ elephant
Time	B-Time	The beginning of a time	غروب/ Sunset
	I-Time	The continuation (inside) of a time	الشمس/ Sun
Month	B-Month	The beginning of a month	رمضان/ Ramadan
	I- Month	The continuation (inside) of a month	
Day	B-Day	The beginning of a day	ثلاثاء/ Tuesday
	I- Day	The continuation (inside) of a day	
Number	B-Num	The beginning of a number	سبعين/ Seventy
	I-Num	The continuation (inside) of a number	وواحد/ One
Other	O	The word is not a named entity (Other)	ربما/ Maybe

4.1.1 Person (Pers)

Persons are used with people's names such as first and middle names, nicknames or last names which can be tagged together, if adjacent. This is equivalent to MUC-6 tag person [1]. Person names are annotated in the corpus as person entities which consist of single or multiple words that refer to a real single person, fictional characters, or religious deities. Furthermore, the alias, nicknames, titles and roles are also annotated as a person NE whenever they refer to persons with no ambiguity. Nicknames are frequently used in classical Arabic. The most frequently used word in NEs that contains more than two words is the word (بن / the son of). This word is usually found between a person's first name and his father's name.

4.1.2 Location (loc)

Location NEs are used to tag names of cities, countries, states, provinces and other places. In classical Arabic, this class annotates all kinds of places and it is equivalent to MUC-6 tag-location [1]. It is also used to tag small places like markets, valleys, cemeteries, mosques and so forth. Furthermore, some places have different names than their old ones which are available in classical Arabic manuscripts. For example (يلملم, yulamlim) is an old name for a place, which is now called (السعدية, alsaedia).

4.1.3 Organization (Org)

Organization NEs are words that refer to organizational structures in various domains like political parties, enterprises, governmental entities, military, and educational domains, and they are equivalent to MUC-6 tag Organization [1]. This class has been widely used in previous work in the Arabic NER in the modern Arabic language. However, organizations rarely occur in classical Arabic.

4.1.4 Measurement (Meas)

Measurement is the task of assigning a numeric value to characteristics of an object or an event that can be compared to other objects or events, such as length, weight, time, etc. In classical Arabic, there are measurement standards which differ from those ones used in modern Arabic. For example, for length, instead of using meters, they use various units such as (ذراع - cubit), (قبضة - a palm-length), (صاع, SAA) - one of the most famous Islamic measurement – around two and a half kilograms), (مد (MOD) – fill two hands).

4.1.5 Money (Mon)

Money NEs are words that refer to absolute monetary quantities. In the MUC-6 tag Money [1], money is any object or record that is accepted as means of payment for goods or services.

4.1.6 Book

A Book is a written or printed text which consists of pages bounded by a cover. The book class refers to the titles of the book in the textual collections. In classical Arabic, this class includes the title of Islamic holy books which are the texts believed by Muslims to be Allah's books to various prophets throughout humanity's history.

4.1.7 Crime

Crime refers to any sort of violation or attacks against the right or interest protected by religion and law, or the commitment of any act which is contrary to justice. In Islam or Shariah law, a crime is defined as an act which could be by a word or an action that violates the conduct established by the Shariah law. The sources of Shariah law are the Holy Quran and Hadith.

4.1.8 Natural Object (NatOb)

Natural objects are entities for living organisms such as animal, vegetable, insect, and mineral. In NLP applications, this class is very

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

important in several domains [41], especially scientific contexts.

4.1.9 Date

A date is a word or expression which contains numeric values that refer to specific dates, months or years on any form of standard calendars. Arabs know the names of the lunar months (Muharram محرم, Safarصفر). Hijri dates are based on events rather than numbers, such as the elephant year. In the era of Caliph Umar, the years needed to be numbered. The Islamic calendar or Hijri calendar (AH), which is a lunar calendar consists of twelve months in a year of 354 days. It is used by many Muslim countries and by any Muslim believer to determine the start of Islamic events such as the fasting month of Ramadan or the pilgrimage to the Muslim holy city (Mecca) and to observe and celebrate Islamic festivals.

4.1.10 Time

Currently, time is being measured by a standard clock with hours, minute and seconds. However, in ancient times, people normally tended to observe the sun and moon positions as signs to measure and determine times which are kinds of time standards used in this corpus.

4.1.11 Month

A month refers to one specific period of time and there are 12 months in the calendar. The month in the Hijri calendar starts at the birth of a new lunar cycle. Practically, this is based on observing the crescent that marks the end of the previous lunar cycle. Some of the months can be 29 or 30 days depending on the observation of the moon and its visibility. On the other hand, some sects follow a tabular Islamic calendar in which odd-numbered and even-numbered months have 29 days and 30 days, respectively. Among these sects are Dawoodi Bohra Muslims and Shia Ismaili Muslims. There are 77 words in the corpus tagged as Month NEs, which takes 0.1% of the NEs in this corpus.

4.1.12 Day

A day is used to mention a specific period of the seven days of the week. The Islamic weekdays begin at sunset. The first day of the week is called (السبت/ Saturday) and the last day is called (الجمعة/ Friday).

4.1.13 Number (Num)

The number is a word or a symbol that refers to specific numeric values or expressions. Arabic symbols for numbers between 0 and 9 are (٠١٢٣٤٥٦٧٨٩). In classical Arabic, numbers are often expressed or written as words rather than numbers or symbols. An example of this is illustrated by (خمسة وعشرون) twenty-five.

4.2 NE types for specific domains

The second one is a specific type for a specific domain. In the CANERCorps, we have seven types which are related to the Classical Arabic and Islamic domain, namely; Allah, prophet, religion, sect, paradise, hell and clan. Table 6 presents types of NE which were detected to match the Islamic domain.

Table 6: Tags definitions for the Islamic Domain

Type	Short	Definition	Example
Allah(God)	B-Allah	The beginning of Allah names	الرحمن/ The Beneficent
	I-Allah	The continuation (inside) of Allah names	الرحيم/ The Merciful
Prophet	B-Prophet	The beginning of a prophet name	رسول/ Messenger of
	I-Prophet	The continuation (inside) of a prophet name	الله/ Allah
Paradise	B-Para	The beginning of a paradise name	جنة/ Paradise
	I-Para	The continuation (inside) of a paradise name	الخلد/ The eternal
Hell	B-Hell	The beginning of a hell name	نار/ The fire
	I-Hell	The continuation (inside) of a hell name	جهنم/ of hell
Religion	B-Rlig	The beginning of a religion name	اليهودية/ Judaism
	I-Rlig	The continuation (inside) of a religion name	والمسيحية/ Christianity
Sect	B-Sect	The beginning of a sect name	الشيعة/ Shia
	I-Sect	The continuation (inside) of a sect name	السنة/ Sunah
Clan	B-Clan	The beginning of a clan name	بني/ Bani
	I-Clan	The continuation (inside) of a clan name	حارث/ Haraith
Other	O	The word is not a named entity (other)	اختار/ Select

4.2.1 Allah

The word “Allah” (God) in Arabic is used by Muslims to refer to God in Islam in accordance with Muslim beliefs. There are 99 names of Allah in Islam [42] and all of them are proper names. The names of Allah in Arabic are called "أسماء الله الحسنى" “Names of God in Islam” and they are the names by which Muslims regard God. These proper names of God are widely used and described in the Holy Quran and Sunnah (prophetic traditions). In this

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

work, all these names were classified under the NE type of "Allah".

4.2.2 Prophet:

From a religious prospective, a prophet is an individual who has been contacted or sent by the divine/supernatural and serving as his messenger and intermediary with humanity. The message that prophets convey to humanity is called prophecy. In this work, the class 'prophet' refers to names of the prophets according to the Muslim beliefs as well as the word (نبي/ Nabi) itself. Moreover, it is linked to all the prophets, such as Christ Jesus and Moses, peace be upon them. It is also considered as the second NE class that is more related to the Islamic domain.

4.2.3 Paradise (Para)

The term 'Paradise' refers to a place of timeless harmony and unlimited resources of welfare. In the Islamic domain, paradise is the eternal afterlife of peace and harmony with the optimum way of living. Muslims believe that those faithful and righteous people who believe in Allah as the almighty God and Muhammad as his final messenger and follow the religious commands will be rewarded with paradise. There are many names which refer to paradise in Islamic traditions such as (طوبى, دار السلام, جنة الخلد).

4.2.4 Hell

Across many cultures and religions, hell is the place of punishment and torment in the hereafter. In the Islamic domain, hell is commonly called the 'النار' fire and there are other names of hell in the Holy Quran and Islamic traditions as well as the names that refer to gates into hell. Among these names in Islam are جحيم / blazing fire, هاوية / the abyss, and سعير / the blaze. Thus, the class 'hell' is labelled for these names in Arabic sentences.

4.2.5 Religions (Rlig)

A religion is a set or collections of beliefs that an individual or a society adheres to. Various religions have symbols, individuals and books that are thought to be holy and sacred. Among the world widespread religions are Islam 'إسلام', Christianity 'مسيحية', Judaism 'يهودية', Buddhism 'بوذية' and Hinduism 'هندوسية'. These words appear in classical Arabic and more frequently in the Islamic domain. Therefore, it would be better to classify individual type in the NE recognition.

4.2.6 Sect:

A sect which is 'طائفة' in Arabic is defined as a division or subclass of a religion or group[43]. The word sect can refer to a specific belief system of a religion or a group. There are various sects within Muslims or religion of Islam. These are often various groups of Muslim people who follow various Muslim scholars or special religious traditions. In the Islamic domain, there are many names of sects which are distributed in different contexts such as, (الشيعة, alshiyea) (الخوارج, alkhawarij).

4.2.7 Clan

A clan refers to a group of people who share the same descent or kinship and sometimes who share a common interest. In the Arabic culture, clans are normally known as small groups of a larger tribe. In this work, the class clan refers to the names of the clan.

5. CORPUS STATISTICS

This corpus contains 258,241 words, 72,108 of them are annotated as NEs, the percentage of the NEs is 28%. Figure 3 shows the details of words where 72,108 considered as NE, 186,133 are considered as other words and the total is 258,241. Furthermore, the number of words in each NEs' class or type was calculated or counted as presented in Table 7 below with the corresponding statistical histogram. It is evident from the statistical histogram below that the highest number of words was scored by person NEs in this corpus, whereas the lowest number of words accounted for organization NEs. As seen in TABLE 6, the highest number of words in Person entity with 39,158 words while the least entity is organization. As mentioned before, the low values are dedicated to the entities, which are not used too much in CA language.

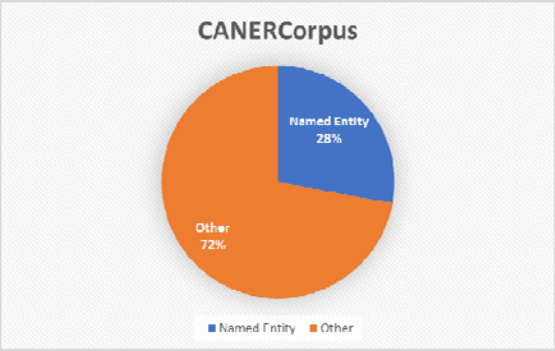

Figure 1 Total number of CANERCorpus

Table 7: Word count and percentage of each NE class

Type	Count	Percentage
Allah	7811	12.95
Prophet	6502	10.77
Pers	39159	64.87
Loc	1349	2.09
Org	9	0.01
Meas	147	0.24
Mon	139	0.23
Book	183	2.24
Date	596	0.95

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

Time	102	0.17
Rlig	184	0.31
Sect	17	0.03
Clan	674	1.11
NatOb	670	1.11
Crime	212	0.35
Para	294	0.49
Hell	245	0.41
Month	77	0.13
Day	31	0.05
Num	13707	1.51
NamedEntity	72108	100.00

As seen in Table 7, the highest number of words in Person entity with 39,158 words while the least entity is organization. As mentioned before, the low values are dedicated to the entities, which are not used too much in CA language. Some of the NEs has low occurrences in CA while others have high occurrences in MSA and still relevant. In hadith, the narrators are many for one Hadith. Therefore, this named entity related to Person gained the highest result among others.

6. DISCUSSION

The corpus that has been collected and annotated in this study was evaluated using the quantitative approach. The evaluation results showed high agreement between among human experts who annotated the named entities in a corpus. For the named entities: Org, Mon, Relig, NatObj, Para, Hell, Month, Day and Num, the Inter-annotator agreement value is 100% per each type. However, there are several issues which make annotators select different categories for some of named entities. One of the common issues in CA is the component word which has more than one category.

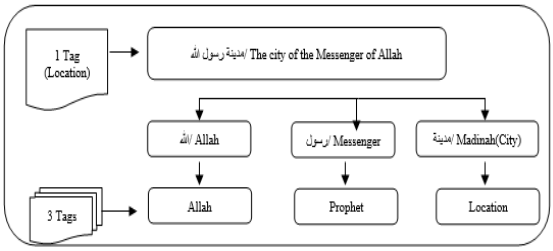

Figure 4: Example of overlapping tags in CANERCorpus

Figure 4, demonstrates how the examples of overlapping tags.

In this example (قادما من مدينة رسول الله, coming from the Prophet's city), one of the annotators considered that in the sentence (مدينة رسول الله, prophet city), there is only one tag, which is a location tag, but another annotator considered that this sentence has three tags, (مدينة, city) which is a location tag, (رسول, prophet), which is another tag related to prophet tag and (الله, Allah), and an Allah tag as well. Based on this, the three annotators were asked to meet, review and discuss their annotations, and to come up with verified and agreed annotations, since problem requires more than just to determining the correct annotation, there is need for discussion in order to arrive at a satisfactory result. Therefore, the annotators agreed to make all depend words as one tag, even if they contain more than one tag, so in this example, the second and third words depend on the first word, so both words function as one word. In other words, we cannot separate such two tags. Moreover, (رسول الله, messenger of Allah), is a prophet tag only, and therefore, we did not separate (رسول, messenger) as a prophet tag and (الله, Allah) as Allah tag.

On the other hand, there are some words that have the same meaning and the same capacity, but are intended to last like a Hell NE type. There are two types: the fire of this world and that is different from the fire of the afterlife that has been linked as one of named entity in this work.

Some adjectives that have been linked to names, such as Prophet peace be upon him. Often in Classical Arabic, the Prophet mentioned is the Messenger of Allah or peace be upon him, without mentioning the name exactly (Muhammed). As per literature, there is scarce researches on CANER and therefore its corpus is unavailable. Thus, NERCorpus which is proposed by this study is new. The need to create a new corpus comes from the nature and requirement of this study since it is for Islamic domain.

7. CONCLUSION AND FUTURE WORK

Corpora have been a major factor in the recent advances in natural language processing development and evaluation. However, CANERCorpus is a free corpus of classical Arabic, manual annotated entities by human experts. It has been constructed with the aim or for the purpose of helping advancing the work on Arabic NE. The corpus and the results we achieved in this study can be used by researchers as gold-standards for classical Arabic and or baselines to test and evaluate their Arabic tools. A gold standard data set is one which contains two different types of domains: a general domain (person, location, organization, measurement, money, book, date, time, natural object, crime, day and number) and a specific domain (Allah, prophet, religion, sect, paradise and hell). However, both domains are related to the classical Arabic which is more pertinent or associated to the Islamic domain. we

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

have extracted approximately 60,362 Arabic NEs that are distributed in different classes from 258,265 tokens. We believe that the resource created is very useful for real world applications in Islamic domain, such as parsing, machine translation and question answering systems.

In the future, we plan to increase the size of our corpus to cover more of the various domains it contains. We also plan to use this corpus to develop morphological analysers and disambiguation systems for CA Arabic.

ACKNOWLEDGMENT

The author would like to express their appreciation to SamaaSoft Company for financial supporting.

REFRENCES:

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," in *COLING*, 1996, pp. 466-471.

[2] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, pp. 3-26, 2007.

[3] K. Shaalan and M. Oudah, "A hybrid approach to Arabic named entity recognition," *Journal of Information Science,* vol. 40, pp. 67-87, 2014.

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

[5] Y. She, X. Zhang, Q. Wang, and Q. Wu, "The potential relationship discovery model based on result fusion for biomedical medicine research," *Journal of Information Science,* vol. 41, pp. 366-382, 2015.

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

[7] R. E. Salah and L. Q. binti Zakaria, "A Comparative Review of Machine Learning for Arabic Named Entity Recognition," *International Journal on Advanced Science, Engineering and Information Technology,* vol. 7, pp. 511-518, 2017.

[8] R. E. Salah and L. Q. binti Zakaria, "Arabic Rule-Based Named Entity Recognition Systems Progress and Challenges," *International Journal on Advanced Science, Engineering and Information Technology,* vol. 7, pp. 815-821, 2017.

[9] Y. Benajiba, M. T. Diab, and P. Rosso, "Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition," *Int. Arab J. Inf. Technol.,* vol. 6, pp. 463-471, 2009.

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

[11] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition: An svm-based approach," in *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*, 2008, pp. 16-18.

[12] N. F. Mohammed and N. Omar, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science,* vol. 8, p. 1285, 2012.

[13] M. Oudah and K. F. Shaalan, "A Pipeline Arabic Named Entity Recognition using a Hybrid Approach," in *COLING*, 2012, pp. 2159-2176.

[14] G. Leech, "Adding linguistic annotation," 2005.

[15] W. Zaghouani, "Critical survey of the freely available Arabic corpora," in *Proceedings of the workshop on free/open-source arabic corpora and corpora processing tools workshop programme*, 2014, p. 1.

[16] K. Bontcheva, L. Derczynski, and I. Roberts, "Crowdsourcing named entity recognition and entity linking corpora," in *Handbook of Linguistic Annotation*, ed: Springer, 2017, pp. 875-892.

[17] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The penn arabic treebank: Building a large-scale annotated arabic corpus," in *NEMLAR conference on Arabic language resources and tools*, 2004, pp. 466-467.

[18] W. Karaa and T. Slimani, "A New Approach for Arabic Named Entity Recognition," *International Arab Journal of Information Technology (IAJIT),* vol. 14, 2017.

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

[20] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," *Journal of Artificial Intelligence Research,* vol. 34, p. 443, 2009.

[21] A. Saif, M. J. Ab Aziz, and N. Omar, "Reducing explicit semantic representation vectors using

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

Latent Dirichlet Allocation," *Knowledge-Based Systems,* vol. 100, pp. 145-159, 2016.

[22] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation," in *HLT-NAACL*, 2007, pp. 196-203.

[23] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *Transactions of the Association for Computational Linguistics,* vol. 2, pp. 231-244, 2014.

[24] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence,* vol. 193, pp. 217-250, 2012.

[25] B. Mohit, N. Schneider, R. Bhowmick, K. Oflazer, and N. A. Smith, "Recall-oriented learning of named entities in Arabic Wikipedia," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 162-173.

[26] M. Attia, A. Toral, L. Tounsi, M. Monachini, and J. van Genabith, "An automatically built named entity lexicon for Arabic," in *n: LREC 2010 - 7th conference on International Language Resources and Evaluation*, Valletta, Malta, 2010, pp. 3614-3621.

[27] F. Alotaibi and M. Lee, "Using Wikipedia as a resource for Arabic named entity recognition," in *Rabat, Morocco. In Proceeding of the 4th International Conference on Arabic Language Processing (CITALA12)*, 2012, pp. 27-34.

[28] M. Azab, H. Bouamor, B. Mohit, and K. Oflazer, "Dudley North visits North London: Learning When to Transliterate to Arabic," in *HLT-NAACL*, 2013, pp. 439-444.

[29] M. Alrabiah, A. Al-Salman, and E. Atwell, "The design and construction of the 50 million words KSUCCA," in *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, 2013, pp. 5-8.

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

[31] N. S. A. Karim and N. R. Hazmi, "Assessing Islamic information quality on the Internet: A case of information about hadith," *Malaysian Journal of Library and Information Science,* vol. 10, p. 51, 2005.

[32] I. Bukhari, "Sahih al-Bukhari," *Kitab Bhavan, New Delhi, India,* 1987.

[33] L. Al-Sulaiti and E. S. Atwell, "The design of a corpus of contemporary Arabic," *International Journal of Corpus Linguistics,* vol. 11, pp. 135-171, 2006.

[34] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research,* vol. 12, pp. 2493-2537, 2011.

[35] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 2001, pp. 1-8.

[36] R. Craggs and M. M. Wood, "Evaluating discourse and dialogue coding schemes," *Computational Linguistics,* vol. 31, pp. 289-296, 2005.

[37] K. Krippendorff, *Content analysis: An introduction to its methodology*: Sage, 2004.

[38] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics,* vol. 34, pp. 555-596, 2008.

[39] S. Sekine, K. Sudo, and C. Nobata, "Extended Named Entity Hierarchy," in *LREC*, 2002.

[40] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 142-147.

[41] C. Whitelaw, A. Kehlenbeck, N. Petrovic, and L. Ungar, "Web-scale named entity recognition," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 123-132.

[42] S. Friedlander, M. Ozak, and H. Amidi, *Ninety-nine names of Allah: the beautiful names*: HarperSanFrancisco, 1993.

[43] M. Sedgwick, "Sects in the Islamic World 1," *Nova Religio: The Journal of Alternative and Emergent Religions,* vol. 3, pp. 195-240, 2000.