International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 2
ISSN: 2088-5334

A Comparative Review of Machine Learning for Arabic Named Entity Recognition

Ramzi Esmail Salah[#1], Lailatul Qadri binti Zakaria[#2]

[#]*Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia.*
E-mail: [1]ramzi@siswa.ukm.edu.my, [2]lailatul.qadri@ukm.edu.my

***Abstract*— Arabic Named Entity Recognition (ANER) systems aim to identify and classify Arabic Named entities (NEs) within Arabic text. Other important tasks in Arabic Natural Language Processing (NLP) depends on ANER such as machine translation, question-answering, information extraction, etc. In general, ANER systems can be classified into three main approaches, namely, rule-based, machine-learning or hybrid systems. In this paper, we focus on research progress in machine-learning (ML) ANER and compare between linguistic resource, entity type, domain, method, and performance. We also highlight the challenges when processing Arabic NEs through ML systems.**

***Keywords*— Arabic named entity recognition; modern standard Arabic; classical Arabic; machine-learning systems**

I. INTRODUCTION

Named Entity Recognition (NER) is very important task in several Arabic Natural Language Processing (NLP) applications. NER can be used for different important tasks, such as Information Extraction (IE), Question Answering (QA), Information Retrieval (IR), and Machine Translation (MT). Applications that employ NER as an important preprocessing step to enhance the overall performance [1]. NER task was firstly introduced at Sixth Message Understanding Conference (MUC-6). However, text can be containing one or more types of names, such as Person, Location, Organization, Sports, and lots of other names from specific domains. These names are called Named Entities (NE). NER seeks to identify and classify these names automatically in text into predefined classes. There has been considerable progress on ANER over the last 10 years [2], and the proposed systems have adapted various NEs methods and techniques which can be roughly classified into rule-based techniques, Machine-Learning (ML) and hybrid approaches. ML approaches are more advantageous as the system can be trained and easily expanded to various language domains [3]. In this paper, we review the work progress in ANER using ML systems. We present a summary of the reported works which include linguistic type, domain, entity type, method, and performance. This paper also discusses the models, and NER features used in ML approaches with some details on the challenges associated with ANER in Arabic text.

A named entity is a term or word that clearly identifies an object from a set of other objects with similar traits. In the expression named entity, the word named limits the scope of entities that have one or many rigid designators that stand for a referent. Usually, rigid designators include proper names, but it depends on the domain of interest that may refer the reference word for the object in the domain as named entities.

A. Arabic Language

There are three forms of the Arabic language: Classical Arabic (CA), Modern Standard Arabic (MSA) and Colloquial Arabic Dialects. CA is the language used in Muslims religious resources, such as Quran and Hadith and in ancient Arabic manuscripts such as poetry. While CA is the foundation of MSA, it has some differences when compared to MSA such as the lexical meaning of words, some grammatical structure, and style. On the other hand, MSA is the current form of Arabic that is considered the official version of Arabic used by governments, agencies, and individuals. The third type of Arabic language is called colloquial Arabic which has mainly used for speaking and exists in various forms depending on the region or country.

B. Arabic Language challenges

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

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

2) Complex Morphology: The complex morphology is common in Arabic text due to Arabic is an inflected language with very rich morphological variations. Various lexical forms can be obtained from different patterns of agglutination. The morphological issue has been handled in several natural language processing applications and tasks such as machine translation [4], noun compound extraction [5], word sense disambiguation [6], semantic relatedness measurement [7], and mapping lexical sources [8]. The experimental results showed that the stemming could improve the traditional NLP applications and tasks.

3) Lack of Resource: Arabic language resources are not adequate, and those available have limited coverage [9]. Some corpora created by individual researchers are available for free to the public [10] while others are available under license agreement [11]. Furthermore, due to recent attention to NER systems for the Arabic language, it is now common to find some Arabic corpora with considerable size available on the web, but many still have limited tools and functions to support Arabic corpus base research.

4) Capitalization Issue: Arabic orthography has no capital letters to distinguish initial letters of proper names like other languages such as those based on Latin-scripted (e.g. English). Thus, the detection of NEs, either expressed in single words or sequence of words, is difficult (Farber et al. 2008). The vagueness created by the disappearance of this element (i.e. capital letters) is further expanded by the way that most Arabic places, proper nouns or things (NEs) are indistinct from words that are common nouns and descriptive words which are non-NEs. Consequently, a methodology depending solely on nouns dictionaries to handle this issue would be uncertain [5]. As an example, the Arabic proper noun اكرم/ (Akram) will serve different meaning in a sentence according to its context; it can be a verb (His/her honoured) or a person name (Akram) and a superlative (the-more-generous).

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

6) Divergence in Writing Style: Arabic language as others do has transcriptional vagueness associated with NEs borrowed from different languages. The problem comes from the variously transliterated ways a word can have [12]. As an example, the word "Google" in English when it is transliterated into Arabic can be written in various spellings using Arabic scripts even though the meaning is still the same. "Uber" can be spelled as أبر or ،أوبر .

II. Material and Method

Machine learning (ML) is the most widely used NER approach in Arabic language and others as well. ML techniques use the features of text and words to recognize NEs. The following two sections summarize the common features used in Arabic ML NER systems and related works on Arabic ML NER systems.

A. Features in NER Used ML Systems

In ANER using ML, there are some features or attributes that are used to recognize NEs. Features in NER are properties or descriptors attributes of words. Feature engineering is a foremost essential task of NER for all classifiers. Word feature can be Features can be specified in numerous ways using one or more Boolean or binary values, numeric or nominal values. The common features are as follow:

1) Word: The word itself, it refers to the distribution of each NE type in the Corpus.

2) Word-Left/Right: Analysis neighbor words (left/Right) of length up to n. The analysis comes in several types, such as part of speech, or Named Entity tags from NER system are used as features.

3) Word Length: This feature can be used to check if the length of a word is less than three or not because it is found that very short words are not named entities [2].

4) Special Marker: This Feature helps for identifying the presence of some special symbols or markers within the text.

5) Word Prefix/ Suffix: The word suffix/prefix feature uses pattern matching to capture word prefix/suffix of length up to n [13]. Also, suffix and prefix Rarely come as NE, the feature could be a good sign for NE existence.

TABLE I
EXAMPLE OF PREFIX AND SUFFIX

Word	Translation	Lemma	Prefix	Suffix
وعلمه	and taught him	علم/ taught	و / and	ـه / him
العربية	Arabic	عرب/ Arabs	ال / the	ية /
رمزي	Ramzi	Ramzi	-	-

6) Capitalization: A binary feature indicating the existence of capitalization information on the gloss corresponding to the Arabic word [2].

7) Lexical Match between Arabic and English: Lexical match between Arabic and English through the use of a bilingual lexicon of morphological analyzer [14]. For example, *Google* may be transliterated to Arabic as (جوجل,) or (غوغل) or (قوقل,). Thus, in the training corpus, if *Google* has only appeared with the first transliteration, the classifier cannot classify the second transliteration. Due to the multiplicity of coiners for Arabic across Arab countries, most untranslatable terms have been transferred in several forms [8] to Arabic. For example, 'biome' and 'pixel' have been transferred to Arabic in different lemmas as ('بيوم' versus 'مجال حيوي بيئي') and ('نقطة ضوئية' versus 'عنصورة') in

Arabic WordNet and Arabic Wikipedia. The bilingual features could improve the accuracy of the semantic similarity between two concepts in different knowledge resources.

8) Nationality Feature: This feature is a merge between two types lexical and a contextual feature. For example, (وصل الرئيس المصري محمد مرسي إلى تركيا, Egyptian president Mohammed Morsi arrived in Turkey). Nationality feature is a binary feature to determine whether the word is recorded in the nationality list or not [1].

9) Trigger Words (Key Words) Feature: One of the important features that guide to identify the NE and can take various forms such as verb list or noun list, it is also called indicator feature. This feature determines if the word is in one of the lexical triggers lists. There are several Arabic terms that have been exploited to identify the named entities in natural language documents Saif, et al. [15] introduced Arabic terms as trigger words for identifying named entity types in Wikipedia articles. These trigger words successfully performed to classify the concepts in Arabic Wikipedia using the category-based technique.

TABLE II
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

10) Blacklist Feature: It is performed using Blacklist dictionaries containing entries which should be rejected as Named Entities. This feature is a twofold feature to determine if the word is in the blacklist. For example, (رئيس الجمهورية القائد الأعلى, the President supreme commander), here the phrase (القائد الأعلى, the supreme commander] is invalid NE.

11) Stop Words Feature: Stop words are frequent words that cannot be part of named entities. This feature is to determine if the word is in the stop words list.

TABLE III
EXAMPLE OF STOP WORDS

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

12) Gazetteer Feature: The gazetteer consists of lists storing specific information such as people's names, organizations names, locations names, days of the week, etc. This feature is to determine if the targeted word exists in any gazetteer class [1].

13) Rule-based Features: These contextual features include the NE type. The NE tags predicted by the rule-based from NER system are used as features.

14) Surrounding Word Feature: Surrounding words that either come before or after a targeted word or token are used a feature to decide if this targeted word is an NE.

15) Infrequent Word: Infrequent words are obtained by calculating the word frequency in the used corpus during the training phase and then selecting the cut-off frequency to build the binary feature.

16) Part-of-Speech (POS) Feature: One of the important feature, often used with ML. This feature identifies the word part of speech class (e.g. verbs, nouns, pronouns, etc.).

17) Syntactic-based Features: Use syntactic rules to label phrases which can be noun or verb phrases.

18) Morphology-based Feature: A group of features extracted from the morphology of the language, it is one of important feature and used widely. The famous one for generation morphology features is MADA [16] they have more than 13 features as shown in Fig. 1.

Feature	Feature value definition
Aspect	*Verb aspect*: Command, Imperfective, Perfective or Not applicable (NA)
Case	*Grammatical case*: Nominative, Accusative, Genitive, NA or Undefined
Gender	*Nominal gender*: Feminine, Masculine or NA
Mood	*Grammatical Mood*: Indicative, Jussive, Subjunctive, NA or Undefined
Number	*Grammatical number*: Singular, Plural, Dual, NA or Undefined
Person	*Person information*: 1st, 2nd, 3rd or NA
State	*Grammatical state*: Indefinite, Definite, Construct/Poss/Idafa, NA or Undefined
Voice	*Verb voice*: Active, Passive, NA or Undefined
Proclitic 3	*Question proclitic*: No proclitic (NP), NA or Interrogative Particle > a
Proclitic 2	*Conjunction proclitic*: NP, NA, Conjunction fa, Response conditional fa, Subordinating conjunction fa, Conjunction wa, Particle wa or Subordinating conjunction wa
Proclitic 1	*Preposition proclitic*: NP, NA, Particle bi, Preposition bi, Preposition ka, Emphatic Particle la, Preposition la, Response conditional la, Jussive li, Preposition li, Future marker sa, Preposition ta, Particle wa, Preposition wa, Preposition fy, Negative particle lA, Negative particle mA, Vocative yA, Vocative wA or Vocative hA
Proclitic 0	*Article proclitic*: NP, NA, Determiner, Negative particle lA, Negative particle mA, Relative pronoun mA or Particle mA
Enclitics	*Pronominal*: No enclitic, NA, 1st person (plural|singular), 2nd person (dual|(feminine (plural|singular))|(masculine (plural|singular))), 3rd person (dual|(feminine (plural|singular))|(masculine (plural|singular))), Vocative particle, Negative particle lA, Interrogative pronoun (ma|mA|man), Relative pronoun (ma|mA|man) or Subordinating conjunction (ma|mA)
POS	*POS definition*: Nouns, Number Words, Proper Nouns, Adjectives, Adverbs, Pronouns, Verbs, Particles, Prepositions, Abbreviations, Punctuation, Conjunctions, Interjections, Digital Numbers or Foreign/Latin

Fig. 1 MADA Morphological features

B. Learning Methods

Machine learning methods are more capable when compared to rule-based approaches because the system can be trained and can work in various domains. In ML NER system, the aim of the NER method is to transform the identification problem into classification one and then use statistical models to tackle the classification problem. In principle, the ML system recognizes and classify NEs into specific NE' class such as locations, persons, organization, etc. [17]. Most recent studies in NE for all major languages including Arabic use a Machine Learning, also called statistical. ML algorithms have been widely used in order to determine NE tagging decisions from annotated texts. The ML approach to the analysis of language works bottom-up by looking for patterns and relationships to model. ML can be divided into three distinguished types: supervised learning, unsupervised learning, and semi-supervised learning. The most commonly published Machine Learning approaches for Named Entity Recognition are Supervised Learning (SL) techniques which represent the NER problem

as a classification task and require the availability of large annotated datasets. Learning methods are more capable when compared to rule-based approaches because the system can be trained and can work in various domains. The followings are common models that are used in ANER ML systems.

C. Supervised Approach

The supervised ML approach is the earliest and widely applied technique in ML systems. Supervised learning aims to train the data on the certain pattern in order to identify it in the test part. This is a useful method in the field of sentiment analysis by train the data about a pattern that may indicate whether the opinion is positive or negative [18]. This approach needs large annotated corpora and among its important statistical models for NER, a lot of works has been done using the following techniques: Conditional Random Fields (CRF), Hidden Markov Model (HMM), Decision Trees (DT), Maximum Entropy Models (ME), Support Vector Machines (SVM), and Artificial Neural Network (ANN). In the following sub-sections, we introduce Arabic named entity recognition using these supervised techniques.

1) Conditional Random Field (CRF)

CRF is a statistical model that is used for data segmentation and labelling in sequence manner [19]. This model involves the use of many random and related features to identify NEs. CRF, as described in [20], is a probabilistic framework used for segmenting and labeling the sequential data. It is a generalization of Hidden Markov Model in which its undirected graph contains nodes to represent the label sequence y corresponding to the sequence x. CRF finds the label which maximizes the conditional probability $p(y|x)$ for a sequence x.

Benajiba and Rosso [21] have used CRF method in replacement of Maximum Entropy in order to improve system performance. The features used in this system are POS tags and Base Phrase Chunks (BPC), gazetteers and nationality. The reported results showed that this system achieved the high accuracy. The general system performance indicators, i.e. recall, precision and F-measure are 72.77%, 86.90%, and 79.21%, respectively. Another work, a simplified feature set has been proposed by Abdul-Hamid and Darwish [13] to be utilized in Arabic NER. They developed a NER system based on CRF to recognize three types of NEs: Person, Location, and Organization. The system considers only surface features and ignores other kinds of features. The system is tested using ANERcorp and ACE2005 dataset. The system performance indicators on ANERcorp for Precision, Recall, and F-measure are 89%, 74%, and 81%, respectively. The results prove that this system is more accurate than the one reported by Benajiba and Rosso [21].

An integrated approach was developed by AbdelRahman, et al. [22] combined two ML systems to handle Arabic NER including pattern recognition using CRF with bootstrapping. The features include word-level features, POS tag, BPC, gazetteers and morphological features. The system can identify various NEs such as Person, Location, Organization, Device, Car, Cell Phone, Date and Time. The F-measures for previous type is 74.06%, 89.09%, 75.01%, 69.47%, 77.52%, 80.95%, 80.63%, 98.52%, 76.99%, and 96.05%. The results show that the system outperforms Ling Pipe NE recognizer when both are applied to ANERcorp dataset.

Bidhend, et al. [23] presented a CRF-based NER system, which is known by the name Noor. The system can extract person names from religious sources. Corpora of ancient religious text called NoorCorp were developed, focusing on 3 corpora based on three Islamic books and jurisprudence sources in Arabic languages. Noor-Gazette, a gazetteer of religious person names, was also developed. The F-measure for the overall system's performance using new historical, Hadith, and jurisprudence corpora was 99.93%, 93.86%, and 75.86%, respectively.

Another work is Impact of Various Features on the Performance of Conditional Random Field-based Arabic Named Entity Recognition by Morsi and Rafea [24], explore the impact of using different feature types on NER results for Modern Standard Arabic text. The system uses CRF based models. They create baseline model to use results for comparison. The dataset was taken from ANERcorp, and extract four types of named entity (person, location, organization and miscellaneous). The best result for the system is a 68.05 F-Measure.

Zirikly and Diab [25] proposed dialectal Arabic NER system using Egyptian colloquial Arabic. Their machine-learning approach uses CRF approach to recognizing persons and locations NEs. They used NER features, namely, lexical with contextual features, gazetteers, distance from specific keywords and Broun clustering. They build an annotated dataset for Egyptian dialect through manually annotating a portion of the dialectal Arabic (DA) data collected and provided by the linguistic data consortium (LDC) from web blogs. The annotated data was chosen from a set of web blogs that are manually identified by LDC as Egyptian dialect. The F-measure obtained for locations and person names are 91.429 and 49.18, respectively. More recently, NEs in social media domain was investigated by Zirikly and Diab [26] who proposed an NER system without the need for gazetteers for DA using supervised machine Learning approach and CRF.

2) Hidden Markov Model (HMM)

HMM is a statistical model that uses Markov process with hidden states. The mathematics of HMM were originally developed by Bikel et al. [27]. Dahan et al. [28] proposed an Arabic NER system based on HMM. The model uses stemming process to address inflection and ambiguity in the Arabic language. The system is fully automated in recognizing Arabic person, organization, and locations NEs. The system was tested using a developed corpus from many sources including France Press agency, *Assabah* newspaper, and Al Hayat newspaper. The performance indicators are precision, recall with 73% and 77% respectively. The obtained F-measure for persons, organization and location NEs are 79%, 67%, and 78% respectively.

3) Decision Tree (DT)

DT was first developed by Sekine et al. [29]. It is a tree-like model which makes decisions at the nodes. A path in the tree represents a sequential of decisions that are following in order to obtain the output at the terminal (tree leafs). ANER

ML system using DT on the criminal domain in MSA was proposed by Al-Shoukry and Omar [30]. Their proposed system can extract NEs of persons, locations, types of crimes, locations, times and date through DTC (Decision Tree classifier) with features extraction. The dataset was collected from online resources. The best obtained F-measure was 81.35%.

4) Maximum Entropy (ME)

ME model predicts the probabilities using the least number of assumptions, different than the applied restrictions. These restrictions are obtained and derived from the training data, which express the relationship between features and outcomes [31].

An ME Arabic Named Entity Recognition system was developed by Benajiba and Rosso [32] who have developed an ANER system, ANERsys 1.0, which uses ME. They used their own developed linguistics resource called ANERcorp (i.e. an annotated corpus) and ANERgazet (i.e. gazetteers). The adapted features are mainly contextual, lexical, together with gazetteers features. The system can recognize various types of NEs, among them, are Person, Location, and Organization. The ANERsys 1.0 system faced difficulties in finding NEs that are have compound structure which composed of more than one token/word; hence [32] come up with ANERsys 2.0, which uses two-level mechanism for NER: 1) identifying the start and the end points of each NE, 2) categorizing the identified NEs. The overall system's performance in terms of Precision, Recall, and F-measure was 70.24%, 62.08%, and 65.91%, respectively.

5) Support Vector Machine (SVM)

SVM is a well-known technique in machine learning which is sometimes called support vector network [33]. SVM is supervised learning method that involves other learning techniques which analyze data for classification and analysis purposes. ANER using SVM was developed by Benajiba et al. [34]. The features used are contextual, lexical, morphological, gazetteers, POS tags and BPC, nationality and the corresponding English capitalization. The system has been evaluated using ACE Corpora and ANERcorp. The best results are achieved when all the features are considered. Furthermore, Y. Benajiba, M. Diab, and P. Rosso [8] studied the sensitivity of various NEs to different types of features, of ACE data sets using the SVM classifier. The best system results in terms of F-measure was 82.71% for ACE 2003, and 76.43% for ACE 2004, and 81.47% for ACE 2005, respectively.

Benajiba et al. [34] have built multiple classifiers for each NE type adopting SVM and CRF approaches. ACE datasets are used in the evaluation process. According to their results, it cannot be stated whether CRF is better than SVM or vice versa in ANER. Each NE type is sensitive to different features, and each feature plays a role in recognizing the NE in different degrees. The best system's overall performance in terms of F-measure was 83.5% for ACE 2003, 76.7% for ACE 2004, and 81.31% for ACE 2005, respectively.

Further studies conducted by Benajiba et al. [35] have confirmed as well the importance of considering language independent and language-specific features in Arabic NER. Benajiba et al. [35] studied the impact of SVM, ME, and CRF models. The reported results in terms of F-measure was 83.34% for ACE 2003, 77.61% for ACE 2004, and 82.02% for ACE 2005, respectively.

Koulali and Meziane [36] developed an ANER using a combined pattern extractor together with SVM classifier that make use of the patterns from POS identified text. The system can cater the NEs types used in the CoNLL conference, and it used a set of dependent and independent features. The system was trained on 90% of the ANERCorp data and tested on the remainder. The system was tested with different using various combinations of features, and the best result of F-measure was 83.20%.

6) Artificial Neural Networks (ANNs)

ANN, one of the important technologies in artificial intelligence, which is considered to be a common approach to machine learning, ANNs are capable of learning, and they need to be trained.

Mohammed and Omar [37] developed a model for the Arabic language to extract Named entity recognition using neural network technique. He uses ANERcorp and other web resources; the system uses two methods to extract 4 types of named entity (person, location, organization and Miscellaneous). The experiment results compared between Decision Tree and Neural Network using the same data. The neural network achieves 92% while decision Tree gained 87% for precision measurement.

D. Semi-Supervised Learning (SSL)

SSL approach is referred to as bootstrapping, which only requires a set of seeds to initiate the learning process. It is the weakly supervised approach, and a set of preliminary learning tasks are used to train the system.

Althobaiti et al. [38] developed Arabic NER system that combines SS approach with distance learning method by training the SS NER classified by the distance learning method. The system extracts person, location and organization NEs in MSA and can be upgraded easily to extract different NEs types. The dataset used are from online NEWS + BBCNEWS and ANERcorp. Table 1 shows the summary of literature review for ML-Base system for the Arabic language.

III. Result and Discussion

In general, as we see in Table 4, ANER using ML systems have received wide attention recently by researchers. The reported studies use various types of the established ML models such as CRF, SVM, ME and HMM with the majority of them based on CRF model. Moreover, most reported works focus on supervised ML methods with few systems that use semi-supervised method whereas the unsupervised method has not been reported for the Arabic language. On the other hand, the common features in other languages are also adapted to ANER ML systems with modifications that arise from the distinct characteristics of Arabic text. These features are roughly based on word-level features, list lookup, word context and linguistic features. Furthermore, most of ML systems are on MSA Arabic, and very few studies are on classical or dialectal Arabic, and most of the studies depend on a single ML model, and there is a need to investigate more on the integration of some models to obtain

better performance results. Additionally, the Arabic language is distinctive, compared to other languages, as it is highly involved with complex morphology and grammars and most of the proposed ANER ML systems use the common features applied elsewhere. Hence, there is a need to come up with new models and features that are well-suited to the nature of Arabic language in order to tremendously enhance the overall performance and capability of ANER ML systems. The unsupervised approaches such as Latent Dirichlet Allocation (LDA) have been utilized for NER in English [39, 40]. Latent Dirichlet Allocation (LDA) is a probabilistic generative model [41-44] of the text documents for semantic representation according to the assumption that states each document is a mixture of topics. It relies on a set of Dirichlet priors that determine how document topic mixtures might be generated on the basis of latent (random) variables. This approach can be adapted to Arabic NER to address the knowledge acquisition in supervised approaches.

TABLE IV
SUMMARY OF LITERATURE REVIEW FOR ML-BASE SYSTEM

Author	Linguistic type	Entity type	Method	Domain	F- measure
Benajiba and Rosso [32]	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
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

IV. CONCLUSION

Over the last decade, the research on ANER has been growing rapidly. Many researchers have developed ML systems for ANER utilizing the established ML models such as CRF, SVM, ME and HMM with the majority of them based on CRF model. Many works focused on supervised ANER ML studies with little attention to semi supervise type whereas the unsupervised approach has not been reported yet. Moreover, most ANER ML systems focus on MSA domain with negligible attention to classical or colloquial Arabic. Furthermore, the studies on ML NER for MSA texts are focusing on few NEs types and even few domains while other domains have rarely been investigated such as criminal records, sports, religion, drugs, etc.

REFERENCES

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

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

[3] G. Talukdar, P. P. Borah, and A. Baruah, "A Survey of Named Entity Recognition in Assamese and other Indian Languages," arXiv preprint arXiv:1407.2918, 2014.

[4] A. Zollmann, A. Venugopal, and S. Vogel, "Bridging the inflection morphology gap for Arabic statistical machine translation," in Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 2006, pp. 201-204.

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

[6] A. Zouaghi, L. Merhbene, and M. Zrigui, "Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation," Artificial Intelligence Review, vol. 38, pp. 257-269, 2012.

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

[8] 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.

[9] L. Abouenour, K. Bouzoubaa, and P. Rosso, "On the evaluation and improvement of Arabic WordNet coverage and usability," Language resources and evaluation, vol. 47, pp. 891-917, 2013.

[10] 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.

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

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

[13] A. Abdul-Hamid and K. Darwish, "Simplified feature set for Arabic named entity recognition," in Proceedings of the 2010 Named Entities Workshop, 2010, pp. 110-115.

[14] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 284-293.

[15] 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.

[16] N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, 2009, pp. 102-109.

[17] A. Mansouri, L. S. Affendey, and A. Mamat, "Named entity recognition approaches," International Journal of Computer Science and Network Security, vol. 8, pp. 339-344, 2008.

[18] M. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," International Journal on Advanced Science, Engineering and Information Technology, vol. 6, pp. 1067-1073, 2016.

[19] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003, pp. 188-191.

[20] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the eighteenth international conference on machine learning, ICML, pp. 282-289.

[21] Y. Benajiba and P. Rosso, "Arabic named entity recognition using conditional random fields," in Proc. of Workshop on HLT & NLP within the Arabic World, LREC, 2008, pp. 143-153.

[22] S. AbdelRahman, M. Elarnaoty, M. Magdy, and A. Fahmy, "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," International Journal of Computer Science Issues (IJCSI), vol. 7, 2010.

[23] M. Bidhend, B. Minaei-Bidgoli, and H. Jouzi, "Extracting person names from ancient Islamic Arabic texts," in Proceedings of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International Conference on Language Resources and Evaluation (LREC 2012), 2012, pp. 1-6.

[24] A. Morsi and A. Rafea, "Studying the impact of various features on the performance of Conditional Random Field-based Arabic Named Entity Recognition," in Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, 2013, pp. 1-5.

[25] A. Zirikly and M. Diab, "Named entity recognition for dialectal arabic," ANLP 2014, p. 78, 2014.

[26] A. Zirikly and M. Diab, "Named entity recognition for arabic social media," in Proceedings of naacl-hlt, 2015, pp. 176-185.

[27] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder," in Proceedings of the fifth conference on Applied natural language processing, 1997, pp. 194-201.

[28] F. Dahan, A. Touir, and H. Mathkour, "First Order Hidden Markov Model for Automatic Arabic Name Entity Recognition," International Journal of Computer Applications, vol. 123, 2015.

[29] S. Sekine, R. Grishman, and H. Shinnou, "A decision tree method for finding and classifying names in Japanese texts," in Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

[30] S. Al-Shoukry and N. Omar, "Proper Nouns Recognition In Arabic Crime Text Using Machine Learning Approach," Journal of Theoretical & Applied Information Technology, vol. 79, 2015.

[31] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in Proc. of the Sixth Workshop on Very Large Corpora, 1998.

[32] Y. Benajiba and P. Rosso, "ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information," in IICAI, 2007, pp. 1814-1823.

[33] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, pp. 273-297, 1995.

[34] 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.

[35] 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.

[36] R. Koulali and A. Meziane, "A contribution to Arabic named entity recognition," in ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on, 2012, pp. 46-52.

[37] N. F. Mohammed and N. Omar, "Arabic named entity recognition using artificial neural network," Journal of Computer Science, vol. 8, p. 1285, 2012.
[38] M. Althobaiti, U. Kruschwitz, and M. Poesio, "Combining Minimally-supervised Methods for Arabic Named Entity Recognition," Transactions of the Association for Computational Linguistics, vol. 3, pp. 243-255, 2015.
[39] G. Xu, S.-H. Yang, and H. Li, "Named entity mining from click-through data using weakly supervised latent dirichlet allocation," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 1365-1374.
[40] I. Bhattacharya and L. Getoor, "A Latent Dirichlet Model for Unsupervised Entity Resolution," in SDM, 2006, p. 59.
[41] M. Andrews and G. Vigliocco, "The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation," Topics in Cognitive Science, vol. 2, pp. 101-113, 2010.
[42] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.
[43] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," Psychological review, vol. 114, pp. 211-244, 2007.
[44] A. Saif, M. J. Ab Aziz, and N. Omar, "Reducing explicit semantic representation vectors using Latent Dirichlet Allocation," Knowledge-Based Systems, vol. 100, pp. 145-159, 2016.