A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach

Mai Mohamed Oudah[1] *Khaled Shaalan*[1,2]
(1) The British University in Dubai, UAE
(2) (Fellow) School of Informatics University of Edinburgh, UK

Oudah.Mai@gmail.com, Khaled.Shaalan@buid.ac.ae

ABSTRACT

Most Arabic Named Entity Recognition (NER) systems have been developed using either of two approaches: a rule-based or Machine Learning (ML) based approach, with their strengths and weaknesses. In this paper, the problem of Arabic NER is tackled through integrating the two approaches together in a pipelined process to create a hybrid system with the aim of enhancing the overall performance of NER tasks. The proposed system is capable of recognizing 11 different types of named entities (NEs): Person, Location, Organization, Date, Time, Price, Measurement, Percent, Phone Number, ISBN and File Name. Extensive experiments are conducted using three different ML classifiers to evaluate the overall performance of the hybrid system. The empirical results indicate that the hybrid approach outperforms both the rule-based and the ML-based approaches. Moreover, our system outperforms the state-of-the-art of Arabic NER in terms of accuracy when applied to ANERcorp dataset, with f-measures 94.4% for Person, 90.1% for Location, and 88.2% for Organization.

KEYWORDS : Natural Language Processing, Named Entity Recognition, Machine Learning.

Title in Arabic

تنسيقٌ متتالٍ في التعرف على أنماط الأسماء العربية من خلال استخدام المنهج الهجين

Abstract in Arabic

تم بناء معظم أنظمة التعرف على أنماط الأسماء العربية من خلال تبني منهجية القواعد أو تبني المنهجية المبنية على تعلم الآلة، بما فيهما من نقاط قوة وضعف. في هذه الورقة، عملية التعرف على أنماط الأسماء في اللغة العربية يتم معالجتها من خلال دمج المنهجيتين معًا في تنسيقٍ متتالٍ لتشكيل المنهج الهجين في محاولة لتحسين أداء مهام التعرف على أنماط الأسماء. النظام المُقترح قادر على التعرف على 11 نوعًا مختلفًا من أنماط الأسماء بما في ذلك أسماء الأشخاص، والأماكن، والمنظمات، والتواريخ، والأوقات، والأسعار (الأموال)، والمقاييس (المقادير القياسية)، والنسب المئوية، وأرقام الهواتف، وردمك (الرقم الدولي المعياري للكتاب)، وأسماء الملفات. وقد تم إجراء تجارب مكثفة باستخدام ثلاث مصنفات مختلفة تُطَبّق تعلم الآلة لتقييم أداء النظام الهجين. تُظهر النتائج التجريبية تفوق المنهج الهجين على كلٍ من المنهج المبني على القواعد والمنهج المبني على تعلم الآلة. يتفوق نظامنا الهجين على أفضل الأنظمة المنشورة في الدوريات العلمية في مجال التعرف على أنماط الأسماء العربية من حيث الدقة عند تطبيق نظامنا على مجموعة البيانات "أنيركورب" بنتيجة معدلات توافقية قدرها: 94.4% في حالة أسماء الأشخاص، 90.1% في حالة أسماء الأماكن، و 88.2% في حالة أسماء المنظمات.

KEYWORDS in Arabic

معالجة اللغات الطبيعية، التعرف على أنماط الأسماء، تعلم الآلة

Proceedings of COLING 2012: Technical Papers, pages 2159–2176,
COLING 2012, Mumbai, December 2012.

1 Introduction

Named Entity Recognition (NER) is the task of detecting and classifying proper names within texts into predefined types, such as Person, Location and Organization names (Nadeau and Sekine, 2007), in addition to the detection of numerical expressions, such as date, time, price and phone number. Machine Translation, Information Retrieval and Question Answering are good examples of Natural Language Processing (NLP) applications that employ NER as an important preprocessing step to enhance the overall performance. In the literature, three types of approaches are used to develop NER systems: rule-based approach, machine learning (ML) based approach and hybrid approach. The rule-based approach relies on handcrafted local grammatical rules, while ML-based approach takes advantage of the ML algorithms that utilize sets of features extracted from datasets annotated with NEs for building NER systems. The hybrid approach combines rule-based approach with ML-based approach together in a pipelined process to improve the overall performance of the system.

Arabic is the official language in the Arab world where more than 300 million people speak Arabic as their native language (Shaalan, 2010). Arabic is a Semitic language and one of the richest natural languages in the world in terms of morphological inflection and derivation. Interest in Arabic NLP has been gaining momentum in the past decade, and some of the tasks have proven to be challenging especially when it comes to Information Extraction due to the language's complex and rich morphology. NER for Arabic has received some attention recently, yet opportunities for improvement in performance are still available. A number of Arabic NER systems have been developed using two types of approaches: the rule-based approach, notably NERA system (Shaalan and Raza, 2008), and the ML-based approach, notably ANERsys 2.0 (Benajiba and Rosso, 2007). Rule-based NER systems rely on handcrafted grammatical rules written by linguists. Therefore, any maintenance applied to rule-based systems is labour-intensive and time consuming especially if linguists with the required knowledge and background are not available. On the other hand, ML-based NER systems utilize ML techniques that require large tagged datasets for training and testing. An advantage of the ML-based NER systems is that they are updatable with minimal time and effort as long as sufficiently large datasets are available. The lack of linguistic resources creates a critical obstacle when it comes to Arabic NLP in general and Arabic NER in particular.

In this paper, the problem of Arabic NER is tackled through integrating the ML-based approach with the rule-based approach to develop a hybrid system in an attempt to enhance the overall performance. To the best of our knowledge, only one recent Arabic NER system (Abdallah, Shaalan and Shoaib, 2012) has adopted the hybrid approach in order to recognize three types of named entities (NEs) including Person, Location and Organization. Abdallah et al. (2012) have used only one ML technique (i.e. Decision Trees) within their system. Our research aims to develop an Arabic hybrid NER system that has the ability to extract 11 different types of NEs including Person, Location, Organization, Date, Time, Price, Measurement, Percent, Phone Number, ISBN and File Name. We extend the ML feature space to include morphological and contextual information. We test three ML algorithms (Decision Trees, Support Vector Machines, and Logistic Regression), and our results show significant performance gains over the state of the art.

The proposed system is composed of two main components: a rule-based component and a ML-based component. The rule-based component is a reproduction of an Arabic rule-based NER

system (Shaalan and Raza, 2008) with modifications and additions in order to enhance the performance. The ML-based component utilizes the ML techniques that have been used successfully in similar NER for other languages to generate a classification model for Arabic NER trained on annotated datasets. The annotated datasets are presented to the ML-based component through a set of features. The feature set is selected to optimize the performance of the ML-based component as much as possible. Two types of linguistic resources are collected and acquired: gazetteers (i.e. predefined lists of NEs or keywords) and corpora (i.e. datasets). Extensive experiments are conducted to evaluate the proposed hybrid system on different dimensions.

The structure of the remainder of this paper is as follows. Section 2 provides some background on NER. Section 3 gives a literature review of NER. Section 4 describes the process followed for data collection. Section 5 illustrates the architecture of the proposed NER system and then describes in details the main components. The evaluation experiments and the results are reported and discussed in Section 6. Finally, a conclusion and proposed future work extension are provided.

2 Background

2.1 NER and NLP Applications

In the 1990s, at the Message Understanding Conferences (MUC) in particular, the task of NER was firstly introduced and given attention by the community of research. Three main NER subtasks were defined at the 6th MUC: ENAMEX (i.e. Person, Location and Organization), TIMEX (i.e. temporal expressions), and NUMEX (i.e. numerical expressions). Customized NER system may require more sub-divisions in one or more of the NER subtasks to fulfil the system goals and objectives, e.g. Location NEs may have sub-types as City, Country, River, Road, etc.

The role of NER within NLP applications differs from one application to another. Examples of NLP applications which find the functionalities of NER useful for their purposes are Information Retrieval, Machine Translation, Question Answering and Text Clustering (Cowie and Wilks, 1996).

- **Information Retrieval (IR).** IR is the task of identifying and retrieving relevant documents out of a database of documents according to an input query (Benajiba, Diab and Rosso, 2009a). There are two possible ways that IR can benefit from NER: 1) recognizing the NEs within the query, 2) recognizing the NEs within the documents to extract the relevant documents taking into consideration their classified NEs. For example, if the input query has the word "مايكروسوفت" maAykruwsuwft[1] "Microsoft", an Organization NE, any documents that include Microsoft is considered relevant and retrieved.
- **Machine Translation (MT).** MT is the task of translating a text into another natural language. NEs need special handling in order to be translated correctly. Hence, the quality of the NE translation component would become an integral part that enhances the performance of the overall MT system (Babych and Hartley, 2003). In the translation from Arabic to Latin languages, such as English, Person names (NEs) can also be found as regular words (non-NEs) in the language without any distinguishing

[1] We used Habash-Soudi-Buckwalter transliteration scheme (Habash, Soudi and Buckwalter, 2007)

orthographic characteristics between the two surface forms. For example, the surface word "وفاء" wafaA' can be used as an adjective that means trustfulness and loyalty, and also as a Person name.

- **Question Answering (QA).** QA application is closely related to IR but with more sophisticated results. A QA system takes questions as input and gives in return concise and precise answers. NER can be exploited in recognizing NEs within the questions to help identifying the relevant documents and then extracting the correct answers (Hamadene, Shaheen and Badawy, 2011; Molla, Zaanen and Smith, 2006). For instance, the NE "الشرق الأوسط" Alšarq AlÂwsaT "Middle East" may be classified as an Organization (i.e. Newspaper) or as a Location according to the context. Hence, the proper classification for the NE will help targeting the relevant group of documents that answer the given query.
- **Text Clustering (TC).** TC may exploit NER in ranking the resulted clusters based on a ratio of entities that is associated with each cluster (Benajiba et al., 2009a). This is reflected in enhancing the process of analyzing the nature of the clusters and also improving the clustering approach in terms of the selected features. For example, Time expressions along with Location NEs can be utilized as factors that give an indication of *when* and *where* the events mentioned in a cluster of documents have happened.

2.2 Arabic Language Characteristics

Applying NLP tasks in general and NER task in particular is very challenging when it comes to Arabic because of its particularities and unique nature. The main characteristics of Arabic that pose non-trivial challenges for NER task are as follows:

- **No Capitalization:** Capitalization is not a feature of Arabic script unlike the European languages where an NE usually begins with a capital letter. Therefore, the usage of the capitalization feature is not an option in Arabic NER. However, the English translation of Arabic words may be exploited in this respect (Farber, Freitag, Habash and Rambow, 2008).
- **The Agglutinative Nature:** Arabic language has a high agglutinative nature in which a word may consist of prefixes, lemma and suffixes in different combination, and that results in a very complicated morphology (AbdelRahman, Elarnaoty, Magdy and Fahmy, 2010).
- **No Short Vowels:** Short vowels, or diacritics, are needed for pronunciation and disambiguation. However, most modern Arabic texts do not include diacritics, and therefore, a word form in Arabic may refer to two or more different words or meanings according to the context they appear, creating a one-to-many ambiguity.
- **Spelling Variants:** In Arabic script, the word may be spelled differently and still refers to the same word with the same meaning, creating a many-to-one ambiguity. For example, the word جرام jrAm 'Gram' can also be written as غرام grAm with the same meaning.
- **Lack of Linguistic Resources:** There is a limitation in the number of available Arabic linguistic resources that are free for research purposes, and many of those available are not suitable for Arabic NER tasks due to the absence of NEs annotations in the datasets or the size of the datasets which may not be sufficiently large. The Arabic gazetteers are rare as well and limited in size. Therefore, researchers tend to build their own Arabic linguistic resources in order to train and evaluate Arabic NER systems.

3 Literature Review

NER revolves around two main goals: 1) the detection of NEs 2) the extraction of those NEs in the form of different predefined types. Three main approaches are used to fulfill those two goals: the rule-based approach, the ML-based approach and the hybrid approach.

3.1 Rule-Based NER

Rule-based NER systems depend on handcrafted linguistic rules to identify NEs within texts using linguistic and contextual clues and indicators (Shaalan and Raza, 2007). Such systems exploit gazetteers/dictionaries as auxiliary clues to the rules. The rules are usually implemented in the form of regular expressions or finite state transducers (Mesfar, 2007). The maintenance of rule-based systems is not a straightforward process since experienced linguists need to be available to provide the system with the proper adjustments (Petasis et al., 2001). Thus, any adjustment to such systems is labour intensive and time consuming.

Maloney and Niv (1998) have presented TAGARAB system which is one of the early attempts to tackle Arabic NER. It is a rule-based system where a pattern matching engine is combined with a morphological tokenizer to recognize Person, Organization, Location, Number and Time. The empirical results show that combining NE finder with a morphological tokenizer outperforms the individual NE finder in terms of accuracy when applied to random datasets from AI-Hayat.

Mesfar (2007) has developed an Arabic component under NooJ linguistic environment to enable Arabic text processing and NER. The component consists of a tokenizer, morphological analyzer and NE finder. The NE finder exploits a set of gazetteers and indicator lists to support rules construction. The system identifies NEs of types: Person, Location, Organization, Currency, and Temporal expressions. The system utilizes the morphological information to extract unclassified proper nouns and thereby enhance the overall performance of the system.

Another work adopting the rule-based approach for NER is the one developed by Shaalan and Raza called PERA (2007). PERA is a grammar-based system which is built for identifying Person names in Arabic scripts with high degree of accuracy. PERA is composed of three components: gazetteers, grammars and filtration mechanism. Whitelists of complete Person names are provided in the gazetteer component in order to extract the matching names regardless of the grammars. Afterwards, the input text is presented to the grammar, which is in the form of regular expressions, to identify the rest of Person NEs. Finally, the filtration mechanism is applied on NEs detected through certain grammatical rules in order to exclude invalid NEs. PERA achieved satisfactory results when applied to the ACE and Treebank Arabic datasets.

As a continuation of Shaalan and Raza (2007) research work, NERA system was introduced in Shaalan and Raza (2008; 2009). NERA is a rule-based system that is capable of recognizing NEs of 10 different types: Person, Location, Organization, Date, Time, ISBN, Price, Measurement, Phone Numbers and Filenames. The implementation of the system was in the FAST ESP framework, where the system has three components as the PERA system with the same functionalities to cover the 10 NE types. The Authors have constructed their own corpora from different resources in order to have a representative number of instances for each NE type.

Elsebai et al. (2009) have proposed a rule-based NER system that integrates pattern matching with morphological analysis to extract Person names from Arabic text. The pattern matching engine utilizes lists of keywords without using predefined lists of Person names. Zaghouani

(2012) has also introduced a rule-based system for Arabic NER (RENAR) to extract Person, Location and Organization NEs. The system is composed of three phases: 1) morphological preprocessing, 2) looking up known NEs and 3) using local grammar to extract unknown NEs. According to the empirical results, RENAR outperforms ANERsys 1.0 (Benajiba et al., 2007), ANERsys 2.0 (Benajiba and Rosso, 2007) and LingPipe[2] in extracting Location NEs when applied to ANERcorp dataset, while LingPipe outperforms RENAR in extracting Person and Organization NEs.

3.2 Machine Learning Based NER

ML-based NER systems take advantage of the ML algorithms in order to learn NE tagging decisions from annotated texts. The most common ML techniques used for NER are Supervised Learning (SL) techniques which represent the NER problem as a classification task and require the availability of large annotated datasets. Among the most common SL techniques utilized for NER are Support Vector Machines (SVM), Conditional Random Fields (CRF), Maximum Entropy (ME), Hidden Markov Models (HMM) and Decision Trees (Nadeau and Sekine, 2007).

Benajiba et al. (2007) have developed an Arabic NER system, ANERsys 1.0, which uses ME. The authors have built their own linguistic resources: ANERcorp (i.e. an annotated corpus) and ANERgazet (i.e. gazetteers). The features used by the system are lexical, contextual and gazetteers features. The system can recognize four types of NEs: Person, Location, Organization and Miscellaneous. The ANERsys 1.0 system used to have difficulties with detecting NEs that are composed of more than one token/word; hence Benajiba and Rosso (2007) developed ANERsys 2.0, which employs a 2-step mechanism for NER: 1) detecting the start and the end points of each NE, 2) classifying the detected NEs. Benajiba and Rosso (2008) have applied CRF instead of ME as an attempt to improve the performance. The feature set used in ANERsys 2.0 was used in the CRF-based system. The features are POS tags and base phrase chunks (BPC), gazetteers and nationality. The CRF-based system achieves higher results in terms of accuracy.

Benajiba et al., (2008a) have developed another NER system based on SVM. 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.

A simplified feature set has been proposed by Abdul-Hamid and Darwish (2010) to be utilized in Arabic NER. They proposed a NER system based on CRF to recognize three types of NEs: Person, Location and Organization. The system considers only surface features (i.e. leading and trailing character n-gram, word position, word length, word unigram probability, the preceding and succeeding words n-gram and character n-gram probability) without taking into consideration any other type of features. The system is evaluated using ANERcorp and ACE2005 dataset. The results show that the system outperforms the CRF-based NER system of Benajiba and Rosso (2008).

Benajiba et al, (2008b) investigated the sensitivity of different NE types to various types of features, i.e. in Benajiba et al., (2008a). They build 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 Arabic NER. Each

[2] LingPipe is available on http://alias-i.com/lingpipe/

NE type is sensitive to different features and each feature plays a role in recognizing the NE in different degrees. Further studies (i.e. Benajiba et al., 2009a; 2009b) have confirmed as well the importance of considering language independent and language specific features in Arabic NER.

AbdelRahman et al. (2010) integrated two ML approaches to handle Arabic NER including CRF and bootstrapping pattern recognition. The feature set used with the CRF classifier includes word-level features, POS tag, BPC, gazetteers and morphological features. The system is developed to extract 10 types of NEs: Person, Location, Organization, Job, Device, Car, Cell Phone, Currency, Date and Time. The results show that the system outperforms LingPipe NE recognizer when both are applied to ANERcorp dataset.

3.3 Hybrid NER

The hybrid approach integrates the rule-based approach with the ML-based approach in order to optimize the overall performance (Petasis et al., 2001). The direction of the processing flow may be from the rule-based system to the ML-based system or vice versa.

To the best of our knowledge, there is only one hybrid NER system for Arabic which has been recently developed by Abdallah, et al. (2012). The hybrid system is capable of identifying Person, Location and Organization NEs. The rule-based component is a re-implementation of the NERA system (Shaalan and Raza 2008) using the GATE tool, while the ML-based component utilizes decision trees to build the NE classifier. Each token/word is represented with a vector of features including the rule-based decisions as a feature. The other features considered are word's length, POS tag, Noun flag (i.e. a binary feature to indicate whether POS tag is Noun or not), gazetteers, statement-end flag, prefix and suffix features. The experimental results show that the hybrid system outperforms the CRF-based NER system built by Benajiba and Rosso (2008) when applied to ANERcorp dataset.

The hybrid NER proves to be feasible and requires further investigations to enhance the scope and improve the overall performance. In this paper, we contribute to hybrid NER for Arabic both in width and depth. We handle the recognition of 11 types of NEs including Person, Location, Organization, Date, Time, Price, Percent, Phone Number, Measurement, ISBN and File Name with high degree of accuracy. We investigate three different ML approaches including Decision Trees (Orphanos, Kalles, Papagelis and Christodoulakis, 1999), SVM (Vapnik, 1995) and Logistic Regression (Hastie, Tibshirani and Friedman, 2009) along with different types of features (including contextual and morphological information) in different combinations to find the feature sets with the optimal performance.

4 Data Collection

Various linguistic resources are necessary in order to develop the proposed Arabic NER system with scope of 11 different categories of NEs. The linguistic resources are of two main categories: corpora and gazetteers. The corpora used in this research are a combination of licensed and free linguistics resources. The licensed linguistics resources[3] are Automatic Content Extraction (ACE) corpora and Arabic Treebank (ATB) Part1 v 2.0 dataset. While the free linguistic resource is: ANERcorp[4] dataset which is freely available for research purposes. In the literature, these

[3] Available for us under license agreement from the Linguistic Data Consortium (LDC)
[4] Available to download on http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

linguistics resources are commonly used for evaluation and comparing with existing systems. We have also built our own corpus for training and evaluating certain types of NEs that were not sufficiently covered, including file names, phone numbers and ISBN numbers. The dataset files have been prepared and annotated using our tag schema and in XML format. Our tag schema includes 11 named entity tags; one for each NE type.

The ACE training datasets covered are Newswire (NW) and Broadcast News (BN). ANERcorp is an annotated dataset built by Yassine Benajiba (Benajiba et al., 2007). Arabic Treebank Part1 v. 2.0 dataset (Maamouri et al., 2003) has no NE annotations and originally designed to support POS tagging in Arabic NLP. Therefore in this research, the ATB dataset has been manually annotated in order to support the Arabic NER task. Our study indicates that the previously listed datasets indicate that they do not include annotation for NEs of types Phone Number, ISBN and File Name. In order to have a dataset with a representative number of NEs of certain types including Phone Number, ISBN and File Name, we acquired our own corpus from different internet resources and did the manual tagging ourselves. The total number of NEs in all datasets (i.e. the number of NE annotations that are used for training and testing purposes) is 23,929 as demonstrated in Table 1.

NE type / Dataset		Per.	Loc.	Org.	Date	Time	Price	Measure.	Percent	Phone No.	File Name	ISBN
ACE 2003	BN	711	1292	493	58	15	17	28	35			
	NW	517	1073	181	20	1	3	14	3			
ACE 2004	BN	1865	3449	1313	357	28	105	51	54			
	NW				67	4	36	30	32			
ACE 2005	BN				154	20	163	60	42			
	NW				37	7	9	22	5			
ANERcorp		3602	4425	2025								
ATB Part1 v 2.0					431	80	168	330	75			
Our own corpus										136	160	126
Total		6695	10239	4012	1124	155	501	535	246	136	160	126

TABLE 1 – The Number of Named Entities in each Reference Dataset

Another type of linguistic resources used is the gazetteers, or dictionaries. The gazetteers for Person, Location and Organization are collected from Shaalan and Raza, (2008), while the gazetteers for the rest of the NE are prepared as part of this research. The total number of NEs/keywords in all gazetteers is 19,328.

5 The System Architecture

The Rule-based and ML-based NER approaches have their own strengths and weaknesses. In this paper, we propose a hybrid architecture that is significantly better than the rule-based or machine-learning systems individually. Figure 1 illustrates the architecture of the hybrid NER system for Arabic. The system consists of two pipelined components: rule-based and ML-based Arabic NER components. The processing goes through three main phases: 1) The rule-based NER phase, 2) The feature engineering phase, i.e. the feature selection and extraction, and 3) the ML-based NER phase.

5.1 The Rule-based Component

The rule-based component in our hybrid system is a reproduction of the NERA system (Shaalan and Raza, 2008) using GATE framework[5]. The rule-based component is built with the capability of recognizing the aforementioned 11 NEs. The percent NE type is introduced in this research and some rules are improved. The rule-based system consists of three main modules: Whitelists (or gazetteers), Grammar Rules (as a set of regular expressions), and a Filtration mechanism (blacklists of invalid NEs).

The GATE environment is used to build the rule-based component. The corpus with its documents is processed using different processing tools and resources such as a tokenizer, gazetteers and grammatical rules. Table 2 illustrates the number of gazetteers and rules implemented within each NE type. The system contains a total of 73 rules and 90 gazetteers.

	Per.	Loc.	Org.	Date	Time	Price	Measure	Percent	Phone No.	File Name	ISBN	Total
# of Gazetteers	11	20	8	12	10	8	7	3	7	3	1	**90**
# of Rules	9	20	9	7	8	4	3	1	7	3	2	**73**

TABLE 2 – The Number of Gazetteers and Rules in each NE Extractor

5.2 The ML-based Component

The ML-based component depends on two main aspects: feature engineering and selection of ML classifiers. The first aspect is the feature engineering which involves the selection and extraction of classification features. The features explored are divided into various categories: rule-based features (i.e. derived from the rule-based component's decisions), morphological features, POS features, Gazetteer features, contextual features, and word-level features. Exploring different types of features and arranging them in sets allow studying the effect of each feature set on the overall performance of the proposed system along different dimensions, including NE type and ML technique.

The second aspect concerns the ML classifier, or function, to be used in the training, testing and prediction phases. Three ML techniques have been explored and examined individually in order to reach a conclusion with regards to the best approach to work with in our hybrid NER system for Arabic. The three techniques are Decision Trees, SVM, and Logistic Regression. The first two techniques were chosen for their high performance in NER in general and Arabic NER in particular; whereas, the third technique is a new investigation that has never been used before in evaluating Arabic NER performance. In this research, WEKA[6], a comprehensive and efficient workbench with support for a large number of ML algorithms, is utilized as the environment of the ML task. The decision tree algorithm is applied using the J48 classifier, SVM with the LibSVM classifier, and Logistic Regression with the Logistic classifier.

[5] Available for free download on http://gate.ac.uk/
[6] The official website of WEKA : www.cs.waikato.ac.nz/ml/weka/

The 11 types of NEs are distributed among three groups according to their nature in which each group has a distinct feature set:

- 1st group: Person, Location and Organization NEs (aka ENAMEX)
- 2nd group: Date, Time, Price, Measurement and Percent NEs (aka TIMEX and NUMEX)
- 3rd group: Phone Number, ISBN and File Name NEs. Notice that the first two types of NE can be considered as NUMEX but they have been moved to this group intentionally because of the nature of their rules and patterns which is specific and limited.

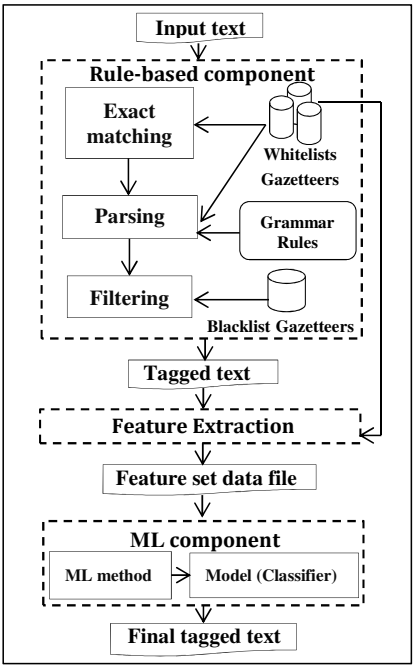

FIGURE 1 – The Architecture of the Hybrid NER System

The three groups of NEs have a generic set of classification features which are common among them, i.e. the following features are used across all three groups.

- *Rule-based features*: The NE type predicted by the rule-based component for the targeted word as well as the NE types for the two immediate left and right neighbors of the candidate word, i.e. NE type for a sliding window of size 5.
- *Morphological Features*: The set of 13 features generated by MADA[7] (Habash and Rambow, 2005).
- *POS tag*: part-of-speech tag of the targeted word estimated by MADA.

[7] MADA is Available for free download on http://www1.ccls.columbia.edu/MADA/MADA_download.html

- *Word length flag*: A binary feature to indicate whether the word length ≥ 3.
- *Dot flag*: A binary feature to indicate whether the word has adjacent dot.
- *Capitalization flag*: A binary feature to indicate the existence of capitalization information on the English gloss (translation) corresponding to the Arabic word.
- *NE type*: NE tag of the word is used along with other features for training the classification model. It is also used as a reference when calculating the accuracy scores. In the prediction phase, this feature (i.e. the NE type itself) is excluded from the selected feature set.

Besides, there are two distinct features that are used in the 1st group:

- *Nominal flag*: A binary feature to indicate whether POS tag is Noun (or Proper Noun).
- *Check Person/Location/Organization Gazetteers feature flags*: A binary feature to indicate whether the word (or left/right neighbour of targeted word) belongs to Person/Location/Organization Gazetteer(s).

Similarly, there are two distinct feature used with the 2nd group:

- *Check POS feature flags*: A binary feature to indicate whether POS tag is Noun_num (i.e. literal number word) (or Proper Noun).
- *Check Date/Time/Price/Measurement/Percent Gazetteers feature flags*: A binary feature to indicate whether the word (or left/right neighbour of targeted word) belongs to Date/Time/Price/Measurement/Percent Gazetteer(s).

Likewise, two distinct features are used with the 3rd group:

- *Nominal flag*: as described in the 1st group feature set.
- *Check Phone Number/ISBN/File Name Gazetteers feature flags*: A binary feature to represent indicate the word (or left/right neighbour of targeted word) belongs to Phone Number/ISBN/File Name Gazetteer(s).

6 Experimental Analysis

6.1 Experimental Setup

We conduct testing and evaluation experiments to test the rule-based component and compare it to the hybrid system. At the level of the hybrid system, experiments are subdivided at three dimensions: the NE type, the ML classifier used, and the inclusion/exclusion of feature groups, with the rule-based decision included as one of the feature groups as will be detailed in the following subsection. Each experiment includes a reference dataset, and an annotated dataset. The reference datasets are the initial datasets described with their tagging details in Section 4 including ACE corpora, ATB part1 v 2.0, ANERcorp and our own corpus. The reference datasets are fed into the rule-based component so that the outputs represent the annotated datasets which are exploited in the feature extraction phase to generate the feature set data files in order to be utilized by the ML-based component.

The performance of the rule-based component is evaluated using GATE built-in evaluation tool, so-called *AnnotationDiff*. This tool enables the comparison of two sets of annotations and the results are presented with the Information Extraction standard measures (i.e. precision, recall and f-measure). On the other hand, the ML approach uses three different functions (or classifiers) to

be applied to the annotated dataset, including decision trees, SVM and logistic regression approaches which are available in WEKA workbench via J48, LibSVM and Logistic classifiers respectively. In this research, 10-fold cross validation is chosen to avoid overfitting. The WEKA tool provides the functionality of applying the conventional k-fold cross-validation for evaluation with each classifier and then having the results represented in the aforementioned standard measures.

6.2 Experiments and Results

A number of experiments have been conducted to evaluate the performance of the proposed hybrid NER system when applied to different datasets in order to extract the various types of NEs applying each of the three different ML techniques. The experiments setting study the performance of the system when the contribution of all features is considered, contribution of pure ML-based features is considered, and after excluding the morphological features generated by MADA (Habash and Rambow, 2005; Roth et al., 2008), i.e. *asp, cas, enc0, gen, mod, num, per, prc0, prc1, prc2, prc3, stt, vox,* and *gloss*. In this way, the following three settings on the level of feature groups are examined:

1. All Features: all features are considered.
2. W/O RB: excluding the rule-based features (pure ML-based mode).
3. W/O MF: excluding the morphological features.

It should be noted that the baseline in all experiments is the performance of the pure rule-based component.

According to the empirical results illustrated in Table 3, the highest performance of our system in terms of Average F-measures when applied on ACE (2003-2004) NW and ANERcorp datasets to extract NEs of the 1st group (i.e. Person, Location and Organization) is achieved by J48 classifier when the 1st feature setting is used, while using J48 classifier with the 3rd setting leads to the highest performance in extracting NEs of the same group from ACE2003 BN dataset.

		ACE2003 NW	**ACE2003 BN**	**ACE2004 NW**	**ANERcorp**
		Avg. F-measure	Avg. F-measure	Avg. F-measure	Avg. F-measure
Rule-based (baseline)		0.6365	0.6087	0.4671	0.6745
J48	All Features	**0.8517**	0.8077	**0.7613**	**0.9090**
	W/O RB	0.8173	0.7633	0.7350	0.8357
	W/O MF	0.8487	**0.8203**	0.7447	0.9047
Libsvm	All Features	0.7953	**0.7653**	**0.7190**	**0.9007**
	W/O RB	0.7453	0.6307	0.6590	0.8100
	W/O MF	**0.7937**	0.7667	0.7117	0.8967
Logistic	All Features	**0.7953**	**0.7693**	**0.7170**	**0.8980**
	W/O RB	0.7577	0.6703	0.6447	0.7753
	W/O MF	0.7827	0.7620	0.7077	0.8857

TABLE 3 – The results of applying the proposed hybrid system on ACE2003 (NW & BN), ACE2004 (NW), and ANERcorp datasets in order to extract NEs of the 1st group

The results illustrated in Table 4 show that the highest performance in terms of Average F-measures when applied on ACE2003 BN, ACE2004 NW & BN, ACE2005 NW & BN and ATB Part1 v 2.0 datasets to extract NEs of the 2nd group (i.e. Date, Time, Price, Measurement and Percent) is achieved by J48 classifier when either the 1st or the 3rd feature setting is utilized, while using Logistic classifier with the 3rd feature setting leads to the highest performance in extracting NEs of the same group from ACE2003 NW dataset. The highest performance of our system in terms of Average F-measures when applied on our own corpus to extract NEs of the 3rd group (i.e. Phone Number, ISBN and File Name) is achieved by either the J48 classifier or the Logistic classifier when the 1st or the 3rd feature setting is utilized as shown in Table 5.

The experimental results show that the adaptation of the hybrid approach leads to the highest performance. It is worth noting that the results of the proposed hybrid system is very close to the results of the rule-based component when it comes to the numerical and temporal expressions, and the two approaches achieve the same results in recognizing NEs of the 3rd group. Therefore, the hybrid approach proves its suitability for the recognition of the three groups of NEs. Also, the decision trees function has proved its comparatively higher efficiency as a classifier in our Arabic hybrid NER system.

		ACE2003 NW	**ACE2003 BN**	**ACE2004 NW**	**ACE2004 BN**	**ACE2005 NW**	**ACE2005 BN**	**ATB**
		Avg. F-measure	Avg. F-measure	Avg. F-measure	Avg. F-measure	Avg. F-measure	Avg. F-measure	Avg. F-measure
Rule-based (baseline)		0.9790	1.0000	0.9766	0.9911	0.9580	0.9839	0.9812
J48	All Features	0.9842	**1.0000**	**0.9874**	**0.9962**	**0.9794**	**0.9870**	**0.9908**
	W/O RB	0.7350	0.6345	0.5742	0.6330	0.6722	0.5703	0.8240
	W/O MF	**0.9864**	**1.0000**	**0.9874**	**0.9962**	**0.9794**	**0.9870**	**0.9908**
Libsvm	All Features	0.9626	0.6985	**0.9818**	0.9090	0.9714	0.9246	**0.9862**
	W/O RB	0.8650	0.2860	0.5840	0.5077	0.8447	0.3610	0.7546
	W/O MF	**0.9672**	**0.7753**	0.9774	**0.9166**	**0.9732**	**0.9344**	**0.9862**
Logistic	All Features	0.9752	0.9280	0.9762	0.9854	0.9702	0.9666	0.9866
	W/O RB	0.6520	0.4123	0.5208	0.5418	0.6642	0.4104	0.7248
	W/O MF	**0.9908**	**0.9334**	**0.9788**	**0.9890**	**0.9774**	**0.9864**	**0.9872**

TABLE 4 – The results of applying our hybrid system on ACE2003, 2004 & 2005 (NW & BN) and ATB Part1 v 2.0 datasets when the 2nd group is the targeted group

		Phone Number	**ISBN**	**File Name**	
		F-measure	F-measure	F-measure	Avg. F-measure
Rule-based (baseline)		1	1	1	1.0000
J48	All Features	**1**	**1**	**1**	**1.0000**
	W/O RB	0.453	0.437	0.899	0.5963
	W/O MF	**1**	**1**	**1**	**1.0000**
Libsvm	All Features	0.996	1	1	0.9987
	W/O RB	0.4	0.148	0.891	0.4797
	W/O MF	0.996	1	1	0.9987
Logistic	All Features	**1**	**1**	**1**	**1.0000**
	W/O RB	0.447	0.518	0.879	0.6147
	W/O MF	**1**	**1**	**1**	**1.0000**

TABLE 5 – The results of applying our hybrid system on our own corpus when the 3rd group is the targeted group

In comparison with the results achieved by ANERsys 1.0 (Benajiba et al., 2007), ANERsys 2.0 (Benajiba and Rosso, 2007), Arabic ML-based NER system using CRF (Benajiba and Rosso, 2008) and the hybrid NER system for Arabic developed by Abdallah et al. (2012) when applied on ANERcorp, our system performs demonstrably better as illustrated by Table 6. As it can be noticed, our hybrid system outperforms the other systems in terms of F-measure in extracting Person, Location and Organization NEs from ANERcorp dataset.

System \ NE Type	Person	Location	Organization
	F-measure	F-measure	F-measure
ANERsys 1.0	0.4669	0.8025	0.3679
ANERsys 2.0	0.5213	0.8671	0.4643
CRF-based system	0.7335	0.8974	0.6576
Abdallah et al. (2012)	0.928	0.8739	0.8612
Our Hybrid System (J48)	**0.944**	**0.901**	**0.882**

TABLE 6 – The results of ANERsys 1.0, ANERsys 2.0, CRF-based system (Benajiba and Rosso, 2008) and Abdallah et al. (2012)'s system compared to our hybrid system's highest performance when applied to ANERcorp dataset

Conclusion and Future Work

The hybrid approach is most recent which integrates rule-based with ML approaches. The integration is more intuitive and linguistically motivated as it conducts an Arabic NER pipeline that combines rule-based features with other features used in machine learning. The proposed hybrid system has achieved an overall improvement of the Arabic NER performance. It is capable of recognizing 11 different types of named entities including Person, Location, Organization, Date, Time, Price, Measurement, Percent, Phone Number, ISBN and File Name. A number of extensive experiments are conducted on three different dimensions including the named entity types, the feature set (divided into groups) and the ML technique to evaluate the performance of our Arabic NER system when applied on different datasets. The experimental results show that the hybrid approach outperforms the pure Rule-based approach and the pure ML-based approach. Our hybrid NER system for Arabic outperforms the state-of-the-art of the Arabic NER in terms of f-measure when applied to ANERcorp dataset with f-measure of 94.4% for Person named entities, f-measure of 90.1% for Location named entities, and f-measure of 88.2% for Organization named entities.

In future work, we intend to enhance the gazetteers and explore the possibility of improving the system with adding more lists. There is also a space for improving the grammatical rules implemented within the rule-based component through analyzing the hybrid system's output in a way to automate the enhancement process. We are also considering the possibility of using different ML techniques other than decision trees, SVM and logistic regression and how this will impact on the overall performance of the system.

Acknowledgments

This research was funded by the British University in Dubai (Grant No. INF004-Using machine learning to improve Arabic named entity recognition).

References

Abdallah, S., Shaalan, K. and Shoaib, M. (2012). Integrating Rule-based System with Classification for Arabic Named Entity Recognition. *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing),* Springer-Verlag, Berlin Heidelberg, pages 311-322.

AbdelRahman, S., Elarnaoty, M., Magdy, M. and Fahmy, A. (2010). Integrated Machine Learning Techniques for Arabic Named Entity Recognition. *International Journal of Computer Science Issues (IJCSI)*, Vol. 7, Issue 4, No 3, pages 27-36.

Abdul-Hamid, A. and Darwish, K. (2010). Simplified Feature Set for Arabic Named Entity Recognition. *Proceedings of the 2010 Named Entities Workshop, (ACL 2010),* pages 110-115.

Babych, B. and Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT (EAMT 2003)*, pages 1-8.

Benajiba, Y., Rosso, P. and Bened´i, J. M. (2007). ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2007)*, Springer-Verlag, Berlin, Heidelberg, pages 143-153.

Benajiba, Y. and Rosso, P. (2007). ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. *Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007)*, pages 1814-1823.

Benajiba, Y. and Rosso, P. (2008). Arabic Named Entity Recognition using Conditional Random Fields. *Proceedings of Workshop on HLT & NLP within the Arabic World (LREC 2008).*

Benajiba, Y., Diab, M. and Rosso, P. (2008a). Arabic Named Entity Recognition: An SVM-Based Approach. *Proceedings of Arab International Conference on Information Technology (ACIT 2008),* pages 16-18.

Benajiba, Y., Diab, M. and Rosso, P. (2008b). Arabic Named Entity Recognition Using Optimized Feature Sets. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 284-293.

Benajiba, Y., Diab, M. and Rosso, P. (2009a). Arabic Named Entity Recognition: A Feature-Driven Study. *IEEE Transactions On Audio, Speech, And Language Processing*, 17(5), pages 926-934.

Benajiba, Y., Diab, M. and Rosso, P. (2009b). Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. *The International Arab Journal of Information Technology,* 6(5), pages 464- 473.

Cowie, J. and Wilks, Y. (1996). Information Extraction. *Communications of the ACM*, 39(1), pages 80-91.

Elsebai, A., Meziane, F. and BelKredim, F. Z. (2009). A Rule Based Persons Names Arabic Extraction System. *Communications of the IBIMA,* pages 53-59.

Farber, B., Freitag, D., Habash, N. and Rambow, O. (2008). Improving NER in Arabic Using a Morphological Tagger. *Proceedings of Workshop on HLT & NLP within the Arabic World (LREC 2008)*, pages 2509- 2514.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573-580.

Habash, N., Soudi, A. and Buckwalter, T. (2007). On Arabic Transliteration. *Arabic Computational Morphology: Knowledge-based and Empirical Methods,* Springer, pages 15-22.

Hamadene, A., Shaheen, M. and Badawy, O. (2011). ARQA: An Intelligent Arabic Question Answering System. *Proceedings of Arabic Language Technology International Conference (ALTIC 2011).*

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2nd ed.). Springer.

Maamouri, M., Bies, A., Jin, H. and Buckwalter, T. (2003). Arabic Treebank: Part 1 v 2.0. *LDC2003T06: Linguistic Data Consortium*, Philadelphia.

Mesfar, S. (2007). Named Entity Recognition for Arabic Using Syntactic Grammars. *Proceedings of the 12th International Conference on Application of Natural Language to Information Systems*, Springer-Verlag, Berlin, Heidelberg, pages 305-316.

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R. (2005). ACE 2004 Multilingual Training Corpus. *Ldc2005t09: Linguistic Data Consortium*, Philadelphia.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L. and Sundheim, B. (2003). Tides Extraction (ACE) 2003 Multilingual Training Data. *Ldc2004t09: Linguistic Data Consortium*, Philadelphia.

Molla, D., Zaanen, M. and Smith, D. (2006). Named Entity Recognition for Question Answering. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006),* pages 51-58.

Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1), pages 3-26.

Orphanos, G., Kalles, D., Papagelis, T. and Christodoulakis, D. (1999). Decision Trees and NLP: A Case Study in POS Tagging. *Proceedings of Annual Conference on Artificial Intelligence (ACAI).*

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V. and Spyropoulos, C. D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. *Proceeding Conference of Association for Computational Linguistics*, pages 426-433.

Roth, R., Rambow, O., Habash, N., Diab, M. and Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking.

Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (HLT-Short '08), pages 117-120.

Shaalan, K. (2010). Rule-based Approach in Arabic Natural Language Processing. *The International Journal on Information and Communication Technologies (IJICT),* 3(3), pages 11-19.

Shaalan, K. and Raza, H. (2007). Person Name Entity Recognition for Arabic. *Proceedings of the 5th Workshop on Important Unresolved Matters*, pages 17-24.

Shaalan, K. and Raza, H. (2008). Arabic Named Entity Recognition from Diverse Text Types. *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*, Springer-Verlag, Berlin, Heidel-berg, pages 440-451.

Shaalan, K. and Raza, H. (2009). NERA: Named Entity Recognition for Arabic. *Journal of the American Society for Information Science and Technology,* 60(8), pages 1652–1663.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). ACE 2005 Multilingual Training Corpus. *Ldc2006t06: Linguistic Data Consortium*, Philadelphia.

Zaghouani, W. (2012). RENAR: A Rule-Based Arabic Named Entity Recognition System. *ACM Transactions on Asian Language Information Processing,* 11(1), Article no. 2, pages 1-13.