Person Name Recognition Using the Hybrid Approach

Mai Oudah[1] and Khaled Shaalan[1,2]

[1] The British University in Dubai, UAE
[2] School of Informatics University of Edinburgh, UK
oudah.mai@gmail.com, khaled.shaalan@buid.ac.ae

Abstract. Arabic Person Name Recognition has been tackled mostly 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 Person Name Recognition 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 Person Name Recognition tasks. 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 Person Name Recognition in terms of accuracy when applied to ANERcorp dataset, with precision 0.949, recall 0.942 and f-measure 0.945.

Keywords: Person Name Recognition, Natural Language Processing, Rule-based Approach, Machine Learning Approach, Hybrid Approach.

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 [19], in addition to the detection of numerical expressions, such as date, time, and phone number. Many Natural Language Processing (NLP) applications employ NER as an important preprocessing step to enhance the overall performance.

Arabic is the official language in the Arab world where more than 300 million people speak Arabic as their native language [22]. Arabic is a Semitic language and one of the richest natural languages in the world in terms of morphology [22]. Interest in Arabic NLP has been gaining momentum in the past decade, and some of the tasks, such as NER, have proven to be challenging due to the language's rich morphology.

Person Name Recognition for Arabic has been receiving increasing attention, yet opportunities for improvement in performance are still available. Most of the Arabic NER systems, which have the capability of recognizing Person names, have been developed using two types of approaches: the rule-based approach, notably NERA system [24], and the ML-based approach, notably ANERsys 2.0 [6]. Arabic rule-based NER systems rely on handcrafted grammatical rules acquired from linguists. Therefore, any maintenance applied to rule-based systems is labor-intensive and time consuming especially if linguists with the required knowledge are not available [21].

E. Métais et al. (Eds.): NLDB 2013, LNCS 7934, pp. 237–248, 2013.

On the contrary, ML-based NER systems utilize learning algorithms that make use of a selected set of features extracted from datasets annotated with named entities (NEs) for building predictive NER classifiers. The main advantages of the ML-based NER systems are that they are updatable with minimal time and effort as long as sufficiently large datasets are available.

In this paper, the problem of Arabic Person Name Recognition 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. Our early hybrid Arabic NER research [1] provided the capability to detect and classify Person NEs in Arabic texts in addition to Location and Organization NEs, where only Decision Trees technique was used within the hybrid system. This technique was applied to a limited set of selected features. The experimental results were promising and assure the quality of the prototype [1]. As a continuation, we extend the ML feature space to include morphological and contextual features. In addition to Decision Trees, we investigate two more ML algorithms: Support Vector Machines and Logistic Regression in the recognition of 11 different types of NEs [20]. In this paper, we report our experience with Arabic Person name recognition in particular. A wider standard datasets are used to evaluate our system. In [20], we reported a set of experimental results which was an indicative of a better system's performance in term of accuracy. Thereafter, more experiments and analysis of results are conducted to assess the quality of the hybrid system by means of standard evaluation metrics.

The structure of the remainder of this paper is as follows. Section 2 provides some background on NER, while Section 3 gives a literature review. Section 4 describes the method followed for data collection. Section 5 illustrates the architecture of the proposed system and then describes in details the main components. The experimental results are reported and discussed in Section 6. Section 7 concludes this paper and gives directions for future work.

2 Background

2.1 NER and NLP Applications

In the 1990s, at the Message Understanding Conferences (MUC), the task of NER was firstly introduced by the research community. 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).

The role of NER within NLP applications differs from an application to another. Examples of those NLP applications (but not limited to) are listed below:

- **Information Retrieval (IR).** IR is the task of identifying and retrieving relevant documents out of a database according to an input query [10]. 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 account their classified NEs. For example, the word "واشنطن" waAšinTun[1] "Washington" can be recognized as a Location NE or a Person NE, hence the correct classification will lead to the extraction of the relevant documents.

- **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 NE translation would become an integral part that enhances the performance of the MT system [4]. In the translation from Arabic to Latin languages, Person names (NEs) can also be found as regular words (non-NEs) in the language without any distinguishing orthographic characteristics between the two surface forms. For example, the surface word "وفاء" wafaA' can be used in Arabic text as a noun which 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 returns 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 [16]. For instance, the words "إرنست ويونغ" Ǎirnist wayuwnγ "Ernst & Young" may be classified as Organization or Person NEs according to the context.

2.2 Arabic Language Characteristics

The main characteristics of Arabic that pose non-trivial challenges for NER are:

- **No Capitalization:** Capitalization is not a feature of Arabic script, unlike Latin languages where NEs usually begins with capital letter. Therefore, the usage of the capitalization feature is not an option in Arabic NER. However, the English translation of Arabic words can be exploited as a feature indicator in this respect [13].
- **The Agglutinative Nature:** Arabic language has a high agglutinative nature in which a word may consist of prefixes, lemma and suffixes in different combination, which results in a very complicated morphology [2].
- **Optional Short Vowels:** In theory, short vowels, or diacritics, are needed for pronunciation and disambiguation. However, practically, most modern standard Arabic texts do not include diacritics, and therefore, a surface form of a word may refer to two or more different meanings according to the context they appear in.
- **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, e.g. the word "جرام" jrAm "Gram" can also be written as "غرام" γrAm.
- **Lack of Linguistic Resources:** There is a limitation in the number of Arabic linguistic resources (corpora (i.e. datasets) and gazetteers (i.e. predefined lists of NEs and keywords)) that are publicly available free for the research purposes. Many of the available corpora are neither annotated with NEs nor include sufficient number of NEs which make them unsuitable for NER task. Therefore, researchers tend to spend tangible efforts to annotate/acquire and verify their own Arabic linguistic resources in order to train and test their systems.

[1] We used Habash-Soudi-Buckwalter transliteration scheme [15].

3 Literature Review of Arabic NER

In this section, we focus on the Arabic NER systems that have the capability to recognize Person names. They are divided to Rule-based and ML-based systems.

3.1 Rule-Based NER

Rule-based NER systems depend on local handcrafted linguistic rules to identify NEs within texts using linguistic and contextual clues, and indicators [24]. 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 [18].

[17] has 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 identify Person, Organization, Location, Number and Time NEs. The empirical results show that combining the NE finder with the morphological tokenizer improves the performance of the system.

[18] has developed an Arabic component under NooJ linguistic environment to enable Arabic NER. 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 morphological information in the recognition of unclassified proper nouns as well.

Another work adopting the rule-based approach for NER is the one developed by [23] called PERA. It is a grammar-based system which is built for identifying Person names in Arabic scripts. PERA is composed of three components: gazetteers, local grammar and filtration mechanism. Whitelists of complete Person names are provided to extract the matching names regardless of the grammars. Afterwards, the input text is presented to the local grammar to identify the rest of Person NEs using the gazetteers. Finally, the filtration mechanism is applied on NEs detected through certain grammatical rules to exclude ambiguous and invalid NEs. PERA achieved satisfactory results when applied to the ACE and Treebank Arabic datasets.

As a continuation of [23] research work, NERA system was introduced in [24, 25]. 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 system was implemented in the FAST ESP framework, where the system has three components as PERA [23] with the same functionalities. The Authors have constructed their own corpora from different resources in order to have a representative number of instances for each NE type.

[12] has proposed a rule-based NER system that integrates pattern matching with morphological analysis to extract Arabic Person names. The pattern matching engine utilizes lists of keywords without using predefined lists of Person names. The performance of the system was compared to PERA [23] despite the fact that PERA is evaluated using different datasets than the ones used for [12]'s system evaluation.

[26] has introduced a rule-based Arabic NER system to extract Person, Location and Organization NEs. The system is composed of three phases: morphological preprocessing, looking up known NEs and using local grammar to extract unknown 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 approach used in ML-based NER is Supervised Learning (SL) approach which represents the NER problem as a classification task. 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 [19].

[5] has developed an Arabic NER system, ANERsys 1.0, which uses ME. The authors have built their own linguistic resources which have become a de facto standard in Arabic NER literature: ANERcorp (i.e. an annotated corpus) and ANERgazet (Person, Location and Organization 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 system raised some difficulties when detecting NEs that are composed of more than one token/word; hence [6] developed ANERsys 2.0, which employs a 2-step mechanism for NER: 1) detecting the start and the end points (boundaries) of each NE, and 2) identifying the NE type. [7] has 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. [8] has 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.

A simplified feature set has been proposed by [3] to be utilized in Arabic NER. They relied on CRF to recognize three types of NEs: Person, Location and Organization. The system considers only surface features without taking into account any other type of features. The system is evaluated using ANERcorp and ACE2005 dataset.

[9] investigated the sensitivity of different NE types to various types of features, i.e. in [8]. They build multiple classifiers for each NE type adopting SVM and CRF approaches. ACE datasets are used in the evaluation process. According to their findings, it cannot be stated whether CRF is better than SVM or the vice versa in Arabic NER. Each NE type is sensitive to different features and each feature plays a role in recognizing the NE in different degrees. Further studies, [10, 11], have confirmed as well the importance of considering language-independent and language-specific features in Arabic NER.

[2] integrated two ML approaches to handle Arabic NER including CRF and bootstrapping pattern recognition. The feature set used 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 system outperforms LingPipe recognizer when both are applied to ANERcorp dataset.

4 Data Collection

The linguistic resources are of two main categories: corpora and gazetteers. The corpora used in this research are Automatic Content Extraction[2] (ACE) corpora and ANERcorp[3] dataset. In the literature, they are commonly used for evaluation as well as comparison with existing systems. The dataset files have been prepared and transformed using our tag schema and in XML format. An example of a Person name in our tag schema is: <Person>هناء</Person>. The three ACE corpora used in this research are ACE 2003 (Newswire (NW) and Broadcast News (BN)) and ACE 2004 (NW) datasets. ANERcorp is an annotated dataset provided by [5]. In this study, the total number of annotated Person NEs covered by all datasets is 6,695 as demonstrated in Table 1. Another type of linguistic resources used is gazetteers. The gazetteers required for Person name recognition are collected as is from [24].

Table 1. The number of Person NEs in each reference dataset

Dataset NE type	ACE 2003 NW	ACE 2003 BN	ACE 2004 NW	ANERcorp	Total
Person	711	517	1865	3602	**6695**

5 The System Architecture

In this article, we propose a hybrid architecture that is demonstrably better than the rule-based or ML-based systems individually. Figure 1 illustrates the architecture of the proposed hybrid system for Arabic. The system consists of two sequential loosely coupled components: 1) a rule-based component that produces NE labels based on lists of NEs/keywords and contextual rules, and 2) an ML-based post-processor intended to make use of rule-based component's NE decisions as features aiming at enhancing the overall performance of the NER task.

5.1 The Rule-Based Component

The rule-based component is a reproduction of the NERA system [24] using the GATE framework[4]. It consists of three main modules: Whitelists (lists of full names), Grammar Rules and a Filtration mechanism (blacklists of invalid names) as illustrated in Figure 1. In GATE, the rule-based component works as a corpus pipeline where a corpus is processed through an Arabic tokenizer, resources including a list of gazetteers, and local grammatical Rules (implemented as finite-state transducers).

[2] Available for us under license agreement from the Linguistic Data Consortium (LDC).

[3] Available to download on `http://www1.ccls.columbia.edu/~ybenajiba/downloads.html`

[4] GATE is freely available at the web link: `http://gate.ac.uk/`

Figure 2 illustrates an example of the Person name rules utilized by the rule-based component. The function of the rule in figure 2 is recognizing expressions that start with "ابو" or "ام" then followed by a First Person Name with the possibility of having a First, Middle or Last Name afterwards. Examples of Person names extracted by this rule: "ابو حسن" (The father of Hassan), and "ام عمر طه" (The mother of Omar Taha).

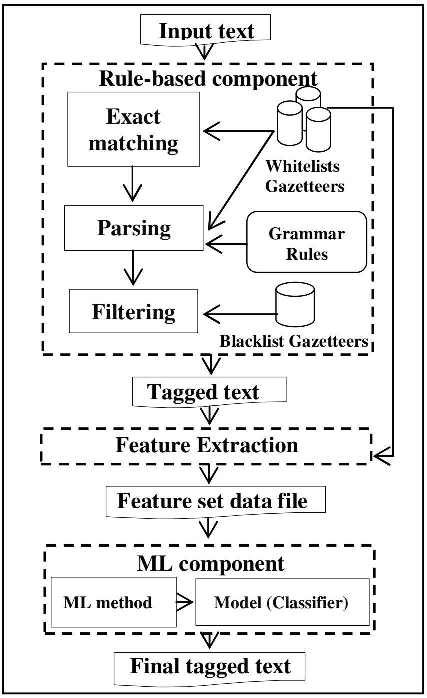

Fig. 1. The Overall Architecture of the Hybrid System

Person Rule in the form of regular expression:

((ابو|ام) + First Name + (First Name | Middle Name | Last Name)?)

Person Rule as implemented in GATE:

```
Rule: PersonRule5
Priority:14
(   ({Token.string == "ابو"}|{Token.string == "ام"})
{Lookup.majorType == "Firsts_v"}
({Lookup.majorType == "Firsts_v"}|{Lookup.majorType == "Middle_vv"}
 |{Lookup.majorType == "Lasts_v"})?     ) :Per
->
:Per.Person={rule="PersonRule1}, :Per.Person={rule="PersonRule5}
```

Fig. 2. An example of Person name rule within the rule-based component

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 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 allow studying the effect of each feature category on the overall performance of the proposed system. The second aspect concerns the ML technique to be used in the training, testing and prediction phases. Three ML techniques (Decision Trees, SVM, and Logistic Regression) have been examined individually to reach a conclusion with regards to the best approach to work in our hybrid system. The first two techniques were chosen for their high performance in NER, while we decided to investigate the effect of the third technique on the proposed system's performance. WEKA[5] is utilized as the environment of the ML task. The classification features used by the ML-based component for Person name recognition are as follows:

- *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:* a set of 13 features generated by MADA [14].
- *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.
- *Check Gazetteers feature flags*: A binary feature to represent whether the word (or left/right neighbour of targeted word) belongs to the Gazetteer set.
- *POS tag:* part-of-speech tag of the targeted word estimated by MADA[6].
- *Nominal flag:* A binary feature to represent whether POS tag is a Noun/Proper Noun.
- Actual NE tag of the word: it is used along with other features for training the classification model. It is also used as a reference for calculating the accuracy.

6 Experimental Results

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 corpora, the ML classifiers, and the

[5] WEKA is available on www.cs.waikato.ac.nz/ml/weka/

[6] MADA is available on: http://www1.ccls.columbia.edu/MADA/

inclusion/exclusion of feature groups. The reference datasets are the initial datasets described with their tagging details in Section 4 including ACE corpora and ANERcorp.

The performance of the rule-based component is evaluated using GATE built-in evaluation tool, so-called *AnnotationDiff*. On the other hand, the ML-based component uses three different functions (or classifiers) to be applied to the datasets, including Decision trees, SVM and Logistic regression approaches which are available in WEKA 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.

6.2 Experiments and Results

A number of experiments have been conducted to evaluate the performance of the proposed system when applied to different datasets. We group similar features together according to the nature of the feature type. We examined six settings of feature groups in order to study their effect on the overall performance: when all features are considered, and when all-but-one feature group are considered. They are:

1. All Features: all features are considered.
2. W/O RB: excluding the rule-based (RB) features (i.e. pure ML-based mode).
3. W/O MF: excluding the morphological features.
4. W/O POS: excluding POS feature.
5. W/O GAZ: excluding Gazetteers features.
6. W/O NbG: excluding neighbors' related features within the Gazetteers features.

The baseline in all experiments is the performance of the pure rule-based component.

Table 2 shows the system's performance in terms of F-measure when applied on ACE2003 (NW & BN), ACE2004 NW, and ANERcorp datasets in order to extract Person NEs. According to the empirical results illustrated in this table, the highest performance of our system when applied on ACE2003 NW and ANERcorp datasets are achieved by J48 classifier when the 6th feature setting is used (i.e. without neighboring features), while using J48 classifier with the 1st setting (i.e. all Features are used) leads to the highest performance when applied on ACE2003 BN and ACE2004 NW datasets.

The experimental results show that the adaptation of the hybrid approach leads to the highest performance. Also, the decision trees function has proved its comparatively higher efficiency as a classifier in our hybrid system. In comparison, the results achieved by [5], [6], [7] and [1] when applied on ANERcorp, have shown that our system performs demonstrably better as illustrated by Table 3. As it can be noticed, our hybrid system outperforms the other systems in extracting Person NEs from ANERcorp dataset. It is worth noting that a comparison between our results and [8, 9]'s results is not possible because their published evaluation lacks sufficient details.

Table 2. The results of applying the proposed hybrid system on ACE2003 (NW & BN), ACE2004 (NW), & ANERcorp datasets in order to extract Person names

		ACE2003 NW	ACE2003 BN	ACE2004 NW	ANERcorp
		F-measure	F-measure	F-measure	F-measure
Rule-based (baseline)		0.7548	0.7646	0.3455	0.6965
J48	All Features	0.932	0.903	0.824	0.944
	W/O RB	0.913	0.886	0.817	0.921
	W/O MF	0.93	0.917	0.812	0.941
	W/O POS	0.924	0.91	0.776	0.94
	W/O GAZ	0.906	0.909	0.785	0.928
	W/O NbG	0.934	0.902	0.82	0.945
Libsvm	All Features	0.919	0.898	0.804	0.942
	W/O RB	0.869	0.844	0.793	0.912
	W/O MF	0.928	0.902	0.805	0.939
	W/O POS	0.919	0.891	0.758	0.939
	W/O GAZ	0.888	0.883	0.753	0.926
	W/O NbG	0.921	0.895	0.795	0.943
Logistic	All Features	0.912	0.902	0.806	0.943
	W/O RB	0.87	0.846	0.799	0.917
	W/O MF	0.903	0.896	0.795	0.935
	W/O POS	0.904	0.896	0.766	0.925
	W/O GAZ	0.889	0.896	0.774	0.919
	W/O NbG	0.906	0.9	0.8	0.937

Table 3. The results of ANERsys 1.0, ANERsys 2.0, CRF-based system [7] and Abdallah et al. [1]'s system compared to our hybrid system's highest performance when applied to ANERcorp dataset in order to extract Person names

System	**Person**		
	Precision	**Recall**	**F-measure**
ANERsys 1.0 [5]	0.5421	0.4101	0.4669
ANERsys 2.0 [6]	0.5627	0.4856	0.5213
CRF-based system [7]	0.8041	0.6742	0.7335
Abdallah et al. [1]	0.949	0.9078	0.928
Our Hybrid System (J48)	**0.949**	**0.942**	**0.945**

7 Conclusion and Future Work

In the literature, the use of either rule-based approach or pure ML-based approach is considered a successful approach for Arabic NER in general and Arabic Person name recognition in particular. Our proposed hybrid approach is distinct from these approaches in that the ML-based subsystem can make use of rule-based decisions determined by the rule-based subsystem in order to improve the performance of Arabic Person name recognition. A number of extensive experiments are conducted on three different dimensions including the dataset, the feature set, and the ML technique used to evaluate the performance of our domain-independent system when applied on a variety of standard datasets. The experimental results prove that the hybrid approach outperforms the pure Rule-based approach and the pure ML-based approach. Moreover, the proposed system outperforms the state-of-the-art of the Arabic Person NER when applied to ANERcorp standard dataset with Precision of 0.949, Recall of 0.942 and F-measure of 0.945 for Person NEs.

In future work, we intend to enhance the gazetteers and explore the possibility of improving the system by adding more lists of predefined Person NEs. There is also a space for improving the local grammar 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 investigating other different ML techniques with our hybrid 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

1. Abdallah, S., Shaalan, K., Shoaib, M.: Integrating Rule-Based System with Classification for Arabic Named Entity Recognition. In: Gelbukh, A. (ed.) CICLing 2012, Part I. LNCS, vol. 7181, pp. 311–322. Springer, Heidelberg (2012)
2. AbdelRahman, S., Elarnaoty, M., Magdy, M., Fahmy, A.: Integrated Machine Learning Techniques for Arabic Named Entity Recognition. IJCSI 7, 27–36 (2010)
3. Abdul-Hamid, A., Darwish, K.: Simplified Feature Set for Arabic Named Entity Recognition. In: Proceedings of the 2010 Named Entities Workshop, pp. 110–115 (2010)
4. Babych, B., Hartley, A.: Improving Machine Translation Quality with Automatic Named Entity Recognition. In: 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), pp. 1–8 (2003)
5. Benajiba, Y., Rosso, P., BenedíRuiz, J.M.: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 143–153. Springer, Heidelberg (2007)
6. Benajiba, Y., Rosso, P.: ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. In: Proceedings of Workshop on Natural Language-Independent Engineering, IICAI 2007, pp. 1814–1823 (2007)
7. Benajiba, Y., Rosso, P.: Arabic Named Entity Recognition using Conditional Random Fields. In: Proceedings of LREC 2008 (2008)
8. Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition: An SVM-Based Approach. In: Proceedings of (ACIT 2008), pp. 16–18 (2008)

9. Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition Using Optimized Feature Sets. In: Proceedings of EMNLP 2008, pp. 284–293 (2008)
10. Benajiba, Y., Diab, M., Rosso, P.: Arabic Named Entity Recognition: A Feature-Driven Study. IEEE Transactions on Audio, Speech and Language Processing 17, 926–934 (2009)
11. Benajiba, Y., Diab, M., Rosso, P.: Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. The International Arab Journal of Information Technology 6, 464–473 (2009)
12. Elsebai, A., Meziane, F., BelKredim, F.Z.: A Rule Based Persons Names Arabic Extraction System. In: Communications of the IBIMA, pp. 53–59 (2009)
13. Farber, B., Freitag, D., Habash, N., Rambow, O.: Improving NER in Arabic Using a Morphological Tagger. In: Proceedings of Workshop on HLT & NLP within the Arabic World (LREC 2008), pp. 2509–2514 (2008)
14. Habash, N., Owen, R., Ryan, R.: 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 (2009)
15. Habash, N., Soudi, A., Buckwalter, T.: On Arabic Transliteration. In: Arabic Computational Morphology: Knowledge-based and Empirical Methods, pp. 15–22 (2007)
16. Hamadene, A., Shaheen, M., Badawy, O.: ARQA: An Intelligent Arabic Question Answering System. In: Proceedings of ALTIC 2011 (2011)
17. Maloney, J., Niv, M.: TAGARAB: A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages (Semitic 1998), pp. 8–15 (1998)
18. Mesfar, S.: Named Entity Recognition for Arabic Using Syntactic Grammars. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) NLDB 2007. LNCS, vol. 4592, pp. 305–316. Springer, Heidelberg (2007)
19. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes 30, 3–26 (2007)
20. Oudah, M.M., Shaalan, K.: A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. In: Proceedings of COLING 2012, pp. 2159–2176 (2012)
21. Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D.: Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In: Proceeding of Association for Computational Linguistics, pp. 426–433 (2001)
22. Shaalan, K.: Rule-based Approach in Arabic Natural Language Processing. IJICT 3, 11–19 (2010)
23. Shaalan, K., Raza, H.: Person Name Entity Recognition for Arabic. In: Proceedings of the 5th Workshop on Important Unresolved Matters, pp. 17–24 (2007)
24. Shaalan, K., Raza, H.: Arabic Named Entity Recognition from Diverse Text Types. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 440–451. Springer, Heidelberg (2008)
25. Shaalan, K., Raza, H.: NERA: Named Entity Recognition for Arabic. Journal of the American Society for Information Science and Technology 60, 1652–1663 (2009)
26. Zaghouani, W.: RENAR: A Rule-Based Arabic Named Entity Recognition System. ACM Transactions on Asian Language Information Processing 11, 1–13 (2012)