Integrating Rule-Based System with Classification for Arabic Named Entity Recognition

Sherief Abdallah[1,2], Khaled Shaalan[1,2], and Muhammad Shoaib[2]

[1] University of Edinburgh, UK
{sherief.abdallah,khaled.shaalan}@buid.ac.ae
[2] British University in Dubai, UAE
shoaibhafeez@hotmail.com

Abstract. Named Entity Recognition (NER) is a subtask of information extraction that seeks to recognize and classify named entities in unstructured text into predefined categories such as the names of persons, organizations, locations, etc. The majority of researchers used machine learning, while few researchers used handcrafted rules to solve the NER problem. We focus here on NER for the Arabic language (NERA), an important language with its own distinct challenges. This paper proposes a simple method for integrating machine learning with rule-based systems and implement this proposal using the state-of-the-art rule-based system for NERA. Experimental evaluation shows that our integrated approach increases the F-measure by 8 to 14% when compared to the original (pure) rule based system and the (pure) machine learning approach, and the improvement is statistically significant for different datasets. More importantly, our system outperforms the state-of-the-art machine-learning system in NERA over a benchmark dataset.

1 Introduction

We propose and implement a simple integration between a (previously developed) rule-based system and a machine-learning classifier for Arabic named entity recognition. A named entity (NE) is a word or a phrase that contains the name of: a person, an organization, or a location among others. For example, the sentence "U.N. official Ekeus heads for Baghdad" contains three named entities: Ekeus is a person, U.N. is an organization and Baghdad is a location [19]. Named entity recognition (NER) is the task of identifying proper nouns in unstructured text. NER is usually an integral component of various Natural Language Processing applications, such as Machine Translation, Search Results clustering, and Question Answering [5]. Most NER approaches can be classified either as a rule-based (RB-NER) or a machine-learning (ML-NER) approach. The RB-NER approach relies on linguistic knowledge, in particular grammar rules, while the ML-NER approach relies on machine learning techniques. RB-NER requires handcrafted rules whereas ML-NER needs an annotated (tagged) corpus. The linguistic knowledge-based approach achieves better results in specific domains, as the gazetteers can be adapted very precisely, and it is able to detect complex entities, as the rules can be tailored to meet nearly any requirement. However, if we deal with an unrestricted domain, it is better to choose the machine learning approach, as it would be expensive (both in terms of cost and time) to acquire and/or derive rules and gazetteers in this case.

A. Gelbukh (Ed.): CICLing 2012, Part I, LNCS 7181, pp. 311–322, 2012.

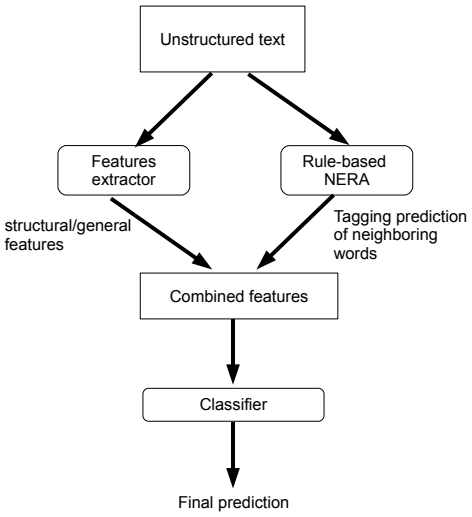

Fig. 1. Block diagram illustrating our proposed integration

The majority of the research on NER focused (naturally) on the English Language with few researchers working on other languages. This paper focuses on NER for the Arabic Language. Arabic is the official language of the ArabWorld (a population 340 million with an explosive growth) and the language of the Quran (the Islamic holy-book, therefore affecting 1.41-1.57 billion Muslims). Arabic is rich in morphology and syntax. Despite the influence of the Arabic language, the research in NER for Arabic is still in its early phases. A major reason for this lag is the lack of available tools (such as taggers and word level analyzers) and linguistic resources (such as named entity tagged corpora and gazetteers). Moreover, the Arabic language is highly challenging to deal with when it comes to perform linguistic grammar based processing. For example, Arabic does not have capital letters; a very important feature in identifying proper nouns. Also it is normally written with optional diacritics (such as short vowels or shadda) which leads to different types of ambiguity in Arabic texts (both structural and lexical), because different diacritics represent different meanings. We describe these challenges in detail in the background section.

We propose in this paper an integration of a rule-based NERA (NER for Arabic) approach, and a machine learning classification approach, as depicted in Figure 1. From the unstructured text, two sets of features are extracted for each word. The first set, which we call the rule-based features, consists of the NE tags predicted by the rule based component for the word in question and a window of surrounding words. The second set of features are general features that are based on our experience.

We verify through extensive experimental results that by complementing the human expertise (through the rule-based component) with automatic fine tuning (through traditional classifiers such as decision trees) we were able to achieve 8-12% improvement over the state-of-the-art NERA system (which used conditional random fields [5]). Interestingly, we also show that relying only on rule-based features does not improve

performance. Also relying on general features does not improve performance (actually leads to a degrading performance). Only when both sets of features are combined does machine-learning classifiers out-performs the-state-of-the-art. These results confirm the value of the integration between RB-NER and ML-NER.

2 Background

In this section we provide the necessary background to understand our contribution. First we give brief overview of Arabic NER, then we describe the rule-based NER for Arabic system which we used as a component in our architecture.

2.1 Arabic Named Entity Recognition

The concept of Name Entity Recognition was born in Message Understanding Conferences in 1990s. In Sixth Message Understanding Conference[1] held in November 1995, the NER task was formally broken down into three subtasks. These subtasks included:

Named Entities - ENAMEX tag. To identify proper names including Person, Organization and Location Names.
e.g. <ENAMEX TYPE="LOCATION">North< /ENAMEX>

Temporal Expression - TIMEX tag. To identify absolute temporal expressions including Date and Time.
e.g. <TIMEX TYPE="DATE">fiscal 1989< /TIMEX>

Number Expression - NUMEX tag. To identify two type of numeric expressions including Money and Percentage.
e.g. <NUMEX TYPE="MONEY">$42.1 million< /NUMEX>

As mentioned earlier, in this work we focus primarily on Arabic NER. The Arabic language has several distinctive challenges when compared to Latin languages [12,17,18]:

Complex Morphology. Arabic is a highly inflected language. Words are formed using stem or root, with prefixes and suffixes characters. This concatenative strategy to form words in Arabic causes data sparseness; hence this peculiarity of the Arabic language poses a great challenge to NER systems [17].

Lack of Capital Letters. Arabic language lacks the capital letters and thus other heuristics have to be applied for detecting Named Entity boundaries such as preceding or succeeding indicator words [17,18].

Non Standard Written Text. The translated and transliterated words to Arabic are not standardized. This is problematic as most of the time all possible spelling variants are not possible to take into consideration [12].

Ambiguity and lack of Diacritization. The written Arabic lacks the Diacritics (short vowels) [2]:
"As most Arabic texts that appear in the media (whether in printed documents or digitalized format) are undiacritized, restoring diacritics is a necessary step for various NLP tasks that require disambiguation or involve speech processing."

[1] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

Missing diacritics are not the only problem. The Arabic words can have different meanings in different contexts which increases the complexity of Named Entity Recognition Systems.

Lack of Resources. The lack of resources for Arabic NER is the major reason of the research in this field being in its infancy. Most of the available resources are either very costly or are of low quality. Thus researchers have to build up their own resources. The lack of using standardized resources thus creates problem of comparing performance among different systems.

We have used the following two corpora for data acquisition and system evaluation (training/testing our classification component):

1. The ACE 2003 Multilingual Training Set[1]
2. ANERcorp Corpus Prepared by Yassine Benajiba[2].

ACE stands for Automatic Content Extraction, a technology that supports automatic processing of human language in textual form.[3] ACE 2003 Multilingual Training Set corpus is distributed by Linguistic Data Consortium (LDC) under the Catalog number LDC2004T09 and ISBN 1-58563-292-9. ACE provides several different files in Standard Generalized Markup Language (SGML) format. These files contain data from Broadcast News and Newswire articles. Each data file in ACE corpus has corresponding XML file which provides Entity information for words in data file. The Entity types covered by ACE 2003 data includes Person, Organization, Location, Facility and Geo Political Entity (GPE). ANERcorp is a corpus prepared by Yassine Benajiba for Named Entity Recognition Task in Arabic Language. With more then 150,000 words annotated for Named Entity Recognition, ANERcorp is ideal for Machine Learning based system as large annotated text is required for better Machine Learning. The details of ANERcorp corpus along with parsing information is described in [8] and [6]. The ANERcorp is easy to parse as each line contains single word with its Entity Information (the corpus is tagged in CONLL format). The possible entity information attached to each tag as described in is listed below:

O Words that are not named entities and referred to as 'Other'.
B-PERS Beginning of Person Name
I-PERS Inside of Person Name
B-ORG Beginning of Organization Name
I-ORG Inside of Organization Name
B-LOC Beginning of Location Name
I-LOC Inside of Location Name
B-MISC Beginning of Miscellaneous Word
I-MISC Inside of Miscellaneous Word

In order to utilize Corpora described in previous sections, we transformed them into XML format using JAVA code. Only Person, Organization and Location entities are

[1] Available to BUID under License.
[2] Available for download from http://users.dsic.upv.es/ ybenajiba/
[3] http://www.itl.nist.gov/iad/mig/tests/ace/

taken into consideration from source corpora during transformation, while other entity types are ignored. For ACE Training set all the files were parsed and transformed into two XML files, one for Broadcast News data and other for Newswire data. All the data of ANERcorp was transformed into single XML file. The XML format is in compliance with the NERA system specification, the rule-based system for Arabic NER that we use in our study. The following section describes the NERA system.

2.2 The NERA System

We have previously developed Named Entity Recognition for Arabic (NERA) prototype. As a proof of concept, we here focus on only three named entities (person name, location, and organization) of those.

We have reimplemented the NERA system [17,18] using the GATE platform.[4] NERA was a rule-based approach for recognizing the most important categories of named entities in Arabic script. The NERA system required a whitelist (gazetteer), a parser and a filtration mechanism. The recognition process included the following two steps: 1) A lookup procedure, called Whitelist, that performed the recognition based on a lookup gazetteer containing lists of known named entities, and 2) A parser, based on a set of grammar rules (represented as regular expressions) derived by analyzing the local lexical context. The Whitelists are fixed static gazetteers (dictionaries) of Named Entities that are matched with target text irrespective of the rules. The exact matches of target text with Whitelist gazetteer entries are reported as Named Entities. Sample entries in gazetteer are shown in Table 1.

Table 1. Sample Data in Gazetteers

Complete Names	حسن نصر الله	محمد سعيد	كوفي أنان
	Hassan Nasar Allah	Muhammad Saeed	Kofi Anan
First Names	عبدالله	عمر	إسحق
	Abdullah	Umar	Ishaq
City Names	مسقط	الطائف	شيكاغو
	Muscat	Taif	Chicago
Prefix Business	الزراعية	التجارية	الصناعية
	Agricultural	Commercial	Industrial

The parser of the NERA system consisted of pattern matching rules that encapsulated linguistic expertise. The rules were based on regular expressions and utilized several different dictionaries within the rules. The Parser was a vital resource as it can deal with peculiarities and complexity of the Arabic language. For instance, the Parser can largely deal with the lack of capitalization for proper nouns by means of using indicator words for named entities. These indicators were used to formulate recognition rules. The NE indicators were obtained as a result of a thorough contextual analysis of various

[4] http://gate.ac.uk/

Arabic scripts. The indicators formed a window around a named entity, which helped in identifying Named Entities without being recognized itself.

3 The Proposed Integrated Approach

Our integration is done by feeding the output of the rule-based system as features to machine-learning classifiers. We call these features the rule-based features. These features are then complemented with other general features that we added through experience. We call the latter features the machine-learning features. We have used Stanford POS Tagger[5] to compute some of these features, such as word category and affixation. All features are then combined and fed to a classifier. We have evaluated several classifier and the results were very similar. We will focus on the decision tree classifier, because its model is easy to understand. Figure 1 illustrates the idea. The features we have used are defined as follows.

Rule-based features. The Named Entity tags from NERA system are used as features. An N-word sliding window (in the experiments we used N=5) is used for each word in corpus. Thus for every word its own tag along with the tag for two left neighbors and two right neighbors are used. Table 2 provides sample instances of these features for 3 words.

Machine-learning features Word-Length. A boolean feature which is TRUE if the word length is greater than three and FALSE otherwise. As pointed out by [10] that very small words are rarely Named Entities.

Noun-Flag. A boolean feature which is TRUE if the part of speech Tag is Noun and FALSE otherwise.

Speech-Tag. Part of speech tag for the current word.

Type-Current. three boolean features, indicating whether the current word is present in the Person Gazetteer, the Organization Gazetteer, or the Location Gazetteer.

Type-Left. similar to Type-Current but for the word to the left of the current word.

Type-Right. similar to Type-Current but for the word to the right of the current word.

Statement-End. A Troolean feature whose value is 1 if the left neighbor of current word is full stop '.', 2 if the right neighbor of current word is full stop '.' and 3 otherwise.

Prefix-Suffix. Prefix of length one for current word, suffix of length one for current word, prefix of length two for current word, and suffix of length two for current word.

4 Experimental Results

Table 3 summarizes the statistical tests of the F-measure that we have conducted, using the J48 decision tree classifier.[6] It is interesting to see that the results are consis-

[5] available at http://nlp.stanford.edu/software/stanford-postagger-2010-05-26.tgz

[6] J48 is an implementation of C4.5 Algorithm for decision trees [16].

Table 2. Sample Rule based features for 5 Word Window

Word	NMinusTwo	NMinusOne	N	NPlusOne	NPlusTwo
الرئيس	OTHER	OTHER	OTHER	OTHER	Person
الروسي	OTHER	OTHER	OTHER	Person	Person
فلاديمير	OTHER	OTHER	Person	Person	OTHER
بوتين	OTHER	Person	Person	OTHER	OTHER

tent across datasets. The rule-based system NERA is at least as good as the Machine-learning approach that uses only rule-based features (MLR) or only our proposed features (ML). However, the Machine-learning approach is significantly better than NERA when all features are used (Hybrid). The results are statistically significant.

Table 3. The F-measure performance using the pure rule-based system (NERA) as a reference point. For example, the first row compares the F-measure performance between NERA and ML approaches. The first column shows that mean difference in F-measure between NERA and ML approaches (F(NERA)-F(ML)) = 6.15, which means that NERA outperforms ML for the first dataset (positive difference). The second column shows the statistical significance (a difference is statistically significant if the two-tail probability is less than 0.05, and lower is better).

	Mean Difference	Two Tail Probability	95% Confidence Interval
ANERcorp Data			
F(NERA) - F(ML)	6.15	0.0011	(-3.2,-9.11)
F(NERA) - F(MLR)	-1.03	0.44956	(-1.91,3.97)
F(NERA) - F(Hybrid)	-8.68	0.000089	(5.75,11.61)
ACE Newswire Data			
F(NERA) - F(ML)	3.14	0.016353	(-5.54,-0.73)
F(NERA) - F(MLR)	2.6	0.137472	(-6.2,1.01)
F(NERA) - F(Hybrid)	-15.48	0.0077	(5.23,25.73)
ACE Broadcast News Data			
F(NERA) - F(ML)	0.7	0.745468	(-5.44,4.04)
F(NERA) - F(MLR)	2.83	0.18167	(-7.26,1.59)
F(NERA) - F(Hybrid)	-6.55	0.0049	(2.55,10.55)

Table 4 compares the results of our integrated approach to previously reported results of the state-of-the-art approach for NERA [7]. We can see that our approach is significantly better for both the person and organization named entities, while our approach has comparable performance in case of the location NE.

An interesting question that we have investigated is why and when our approach disagrees with the RB-NER. To answer this question we investigate here in more depth the resulting decision tree. An example tree that we have obtained from the J48 classifier [16] with all the features described earlier and when applied on ANERcorp Data [8] consists of 1126 leaves and the size of the tree is 1684 nodes in total. Figure 2 shows

Table 4. Comparison of F-measure performance between our proposed hybrid approach and the conditional random fields approach

	Person			Organization			Location			Mean
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**F**
Our integrated approach	94.9	90.78	**92.8**	86.26	85.99	**86.12**	90.6	84.4	**87.39**	88.77
Conditional random fields	80.41	67.42	73.35	84.23	53.94	65.76	93.03	86.67	89.74	76.28

the subtree where top node N (the type predicted by the RB-NER) has the value Organization. This subtree is interesting because it shows cases of disagreements, where the final class in some cases is Location. Consider an example where word ألمانيا (Germany) is shown with three tags and few words surrounding to the left and right of ألمانيا in the ANERcorp Dataset:

فرانكفورت (د ب أ) أعلن اتحاد صناعة السيارات في

Location_ Location_ Organization_ ألمانيا

امس الاول أن شركات صناعة السيارات في ألمانيا تواجه ....

Translation: **"Frankfurt, Auto Industry Association in Germany said the day before yesterday that Automakers in Germany is facing ..."**

In this example the word ألمانيا is followed by first tag "Organization" which is recognized by rule based system. "Location" is the second tag for word ألمانيا and is identified by Decision Tree. The final tag is actual tag in corpus for word ألمانيا and it is also "Location". As per the actual tagging in corpus i.e. "Location", the recognition of word ألمانيا as "Organization" is incorrect by rule based system. The order of tree traversal is given in Figure 2 to correctly classify the word as "Location". The values of the features (used in Decision Tree) for this word are N=Organization, isLookupOrganization = FALSE, NPlusOne = OTHER, Prefix=2, NMinusOne = Organization, Actual=Location. Another similar example is given below:

تحقيق السلام في دارفور . الظواهري قال إن حكومة

Location_ Location_ Organization_ الخرطوم

عاجزة عن حل أزمة دارفور (رويترز) وكان ...

Translation: **"Achieving peace in Darfur. Al-Zawahiri said that the Khartoum government is powerless to solve the Darrfor crisis (Reuters) and was ..."**

In this example also the recognition of the word الخرطوم (Khartoum) by rule based system as "Organization" is incorrect as it is tagged as "Location" in reference corpus. The correct classification of the word الخرطوم is given by Decision tree as "Location". The the order of tree traversal for is given in Figure 2 to correctly classify this word as "Location". The values of features (used in Decision Tree) for this word are N = Organization, isLookupOrganization = FALSE, NPlusOne = OTHER, Prefix = 2, NMinusOne = Organization, Actual=Location.

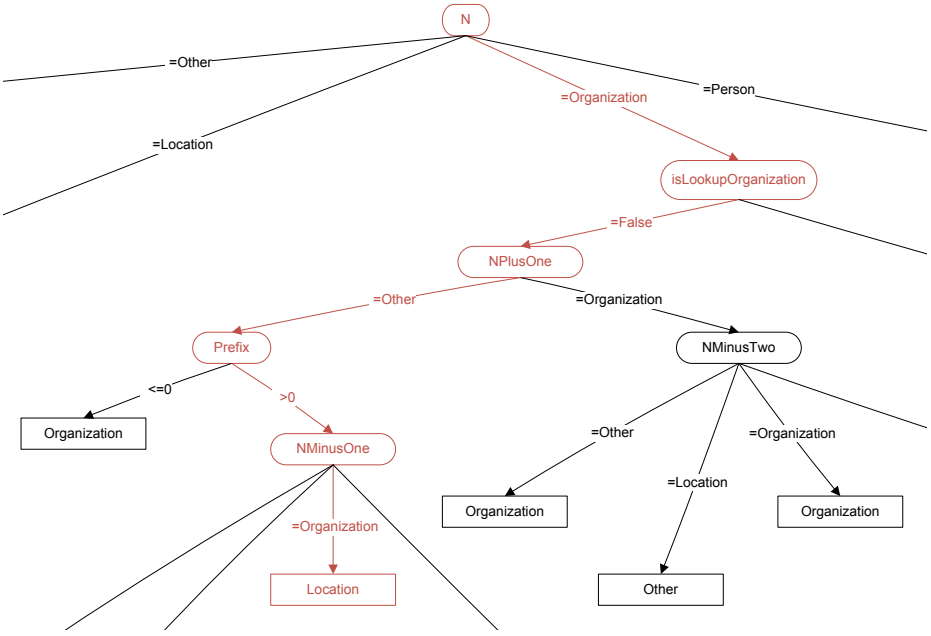

Fig. 2. Part of the decision tree learned by the J48 algorithm for the ANERCorp using all the features. Highlighted path corresponds to the example in the text.

We were initially surprised that the rule-based system was not able to correctly recognize the above Location NEs. However, upon further investigation we have learned that the errors of NERA (the rule-based system) above are actually corpus specific. In the original corpus where NERA rules were developed, an Organization NE would include the organization's location. However, in the ANERcorp corpus an organization location is considered a separate Location NE. Our hybrid approach was successfully able to adapt the rules to account for these differences across corpora.

5 Related Work

As we have mentioned earlier, the work in NER generally fell under either rule-based or machine-learning approaches. Rule based systems allowed expert linguists to handcraft rules for the NER task. This encoding of human expertise required extensive work from expert linguists and usually targeted a single language. As a result, only few researchers used rule-based systems to tackle NER for Arabic. The rules were implemented as regular expression for pattern matching mostly in conjunction with list of lookup gazetteers.

TAGARAB [13] was one of the early systems that used rule based pattern matching for NER in Arabic. TAGARAB used morphological analysis of text in conjunction with pattern matching to achieve higher accuracy as compared to simple pattern matching. Another work discussed the application of local grammar based approach in domain

of Arabic language [20]. The grammar was extracted by applying corpus analysis over range of untagged Arabic corpora. The result was a finite state automata to extract named entities from Arabic text.

Machine Learning is the mostly applied method for NER for all major languages including Arabic. NER was viewed as classification problem, where text features are used to classify words either as a particular NE or as normal text. The features include both language specific features (e.g. Part of Speech information, Morphological features etc) and language independent features (e.g. length of the word etc). A major shortcoming of the machine learning approach is requiring large corpora of annotated text. This shortcoming is more magnified in Arabic NER due to the lack of linguistic resources. Our integrated approach complements this limitation with the human expertise encapsulated in the rule-based component.

One of the early attempts to utilize Machine Learning for NER [3] used word-level features, dictionary Look-Up, part-of-speech tags and punctuation. The reported accuracy of the system was comparable to state-of-the-art rule-based system at the time. ANERSys, an NER System, was based on Maximum Entropy [8]. The baseline results was acquired by assigning each word in the test set a class that was most frequently assigned to it in the training set. Later the training and testing were done using the Maximum Entropy approach. The authors reported significant improvement over baseline results.

The work of ANERSys was extended to ANERSys 2.0 [6]. The approach used Maximum Entropy along with part of the speech information. The same baseline was used as in ANERSys. The authors [6] reported significant improvement over the baseline results, results from ANERSys and results from demo version of Siraj (Sakhr) which is a commercial system for Named Entity Recognition. Using Conditional Random Fields instead of Maximum Entropy for ANERSys system resulted in further improvement [7]. A similar approach used leading and trailing character n-grams in words as features [1], which reported better performance over previous work. Support vector machines with very large number of features were also used [14,11,4,5] but suffered from (very) slow training time and could not incorporate human knowledge if available.

Hybrid approaches combined hand crafted rule based system and Machine Learning system (our approach falls in this category). A recent hybrid approach applied Maximum Entropy (ME) with Hidden Markov Model (HMM) followed by rules to detect NE [9]. Our approach, on the other hand, uses RB-NER component followed by ML-NER. Perhaps the most similar work to our approach used rule-based systems to provide training labels [15]. In the first stage, authors passed text through the rule-based system to tag the words. The tags were then used as the ground truth for training a classifier. In other words, they did not use previously-labeled corpus. In the evaluation stage, text was tagged, independently, by both the Rule Based System and the classifier. Cases of disagreement were presented to an expert Linguist. Unlike our approach, there was no expert-tagged corpus that was utilized in the training phase of Machine Learning Model.

6 Conclusion

We have proposed in this paper an architecture for Arabic Named Entity Recognition that integrates rule-based with machine learning classifiers. As a proof of concept, we

have re-implemented a rule-based system and then integrated it with a decision-tree classifier. Experimental results confirm that our hybrid approach is significantly better than the pure rule-based system or the pure machine-learning classifier. Our approach is also better than the state-of-the-art Arabic NER (which relied on conditional random fields). Our hybrid approach was successfully able to adapt the rules to account for the tagging differences across corpora.

References

1. Abdul Hamid, A., Darwish, K.: Simplified feature set for arabic named entity recognition. In: Proceedings of the 2010 Named Entities Workshop, pp. 110–115. Association for Computational Linguistics, Uppsala (2010), http://www.aclweb.org/anthology/W10-2417
2. Attia, M., Toral, A., Tounsi, L., Monachini, M., van Genabith, J.: An automatically built named entity lexicon for arabic. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta (May 2010)
3. Baluja, S., Mittal, V.O., Sukthankar, R.: Applying machine learning for high performance named-entity extraction. Computational Intelligence 16(4), 586–595 (2000)
4. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: An svm-based approach. In: The International Arab Conference on Information Technology, ACIT 2008 (2008)
5. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition using optimized feature sets. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 284–293. Association for Computational Linguistics, Morristown (2008)
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: IICAI, pp. 1814–1823 (2007)
7. Benajiba, Y., Rosso, P.: Arabic named entity recognition using conditional random fields. In: Workshop on HLT & NLP within the Arabic World. Arabic Language and Local Languages Processing: Status Updates and Prospects (2008)
8. Benajiba, Y., Rosso, P., Benedí Ruiz, 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)
9. Biswas, S., Mishra, S.P., Acharya, S., Mohanty, S.: A hybrid oriya named entity recognition system: Harnessing the power of rule. International Journal of Artificial Intelligence and Expert Systems (IJAE) 1, 1–6 (2010)
10. Ekbal, A., Bandyopadhyay, S.: Voted ner system using appropriate unlabeled data. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, NEWS 2009, pp. 202–210. Association for Computational Linguistics, Morristown (2009)
11. Ekbal, A., Bandyopadhyay, S.: Named entity recognition using support vector machine: A language independent approach. International Journal of Electrical, Computer, and Systems Engineering 4(2), 155–170 (2010)
12. Habash, N.Y.: Introduction to Arabic Natural Language Processing. Mogran & Claypool Publisher (2010)
13. 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. Association for Computational Linguistics, Morristown (1998)

14. Mayfield, J., McNamee, P., Piatko, C.: Named entity recognition using hundreds of thousands of features. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, vol. 4, pp. 184–187. Association for Computational Linguistics, Morristown (2003), http://dx.doi.org/10.3115/1119176.1119205
15. 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: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL 2001, pp. 426–433. Association for Computational Linguistics, Morristown (2001)
16. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
17. 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)
18. Shaalan, K., Raza, H.: NERA: Named entity recognition for arabic. Journal of the American Society for Information Science and Technology, 1652–1663 (2009)
19. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, CONLL 2003, vol. 4, pp. 142–147. Association for Computational Linguistics, Stroudsburg (2003), http://dx.doi.org/10.3115/1119176.1119195
20. Traboulsi, H.: Arabic named entity extraction: A local grammar-based approach. In: Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 139–143 (2009)