International Arab Conference on Information Technology (ACIT'2016)

Exploiting Multilingual Wikipedia to improve Arabic Named Entity Resources

Mariam Biltawi, Arafat Awajan, Sara Tedmori, and Akram Al-Kouz
King Hussein Faculty of Computing Sciences
Princess Sumaya University for Technology
Amman, Jordan
maryam@psut.edu.jo, awajan@psut.edu.jo, s.tedmori@psut.edu.jo, akram@psut.edu.jo

***Abstract*—This paper focuses on the creation of Arabic named entity gazetteer lists, by exploiting Wikipedia and using the Naïve Bayes classifier to classify the named entities into the three main categories: person, location, and organization. The process of building the gazetteer starts with automatically creating the training and testing corpora. The training corpus consists of Arabic text; whereas, the testing corpus is derived from an English text using the Stanford name entity recognition. A Wikipedia title existence check of these English name entities is then performed. Next, if the named entity exists as a Wikipedia page title, a check for Arabic parallel pages is conducted. Finally, the Naïve Bayes classifier is applied to verify or assign new name entity tag to the Arabic name entity. Due to the lack of available resources, the proposed system is evaluated manually by calculating accuracy, recall, and precision. Results show an accuracy of 53%.**

Keywords—Arabic name entity resources; naïve Bayes classifier; Wikipedia

I. INTRODUCTION (*HEADING 1*)

Named Entity Recognition (NER) is a subtask of information extraction, and refers to the process of extracting and classifying some text elements into various pre-defined classes such as names of persons, organizations, locations, date and time expressions, percentages, quantities, and monetary values [1]. Additional classes can include: biological species, genes, proteins, diseases, and anatomy [2]. NER can be useful in many applications such as information retrieval, question answering, machine translation, text clustering, and navigation systems.

Research works focusing on recognizing Named Entities (NEs) from different languages are available but mostly for English. Work on the Arabic NER is still limited due to specific features and challenges of Arabic language. Firstly, there are three types of Arabic language, classical Arabic, Modern Standard Arabic (MSA) and colloquial Arabic. Secondly, Arabic language is generally ambiguous. It is a highly agglutinative language usually written with the short vowels omitted. Moreover, Arabic lacks capitalization, uniformity in writing styles, and resources [3].

To date, three main general purpose tag sets have been devised for Arabic language NE tagging. The first tag set, which consists of three tag elements, was initiated in 1995 during the 6th Message Understanding Conference (MAC-6) [4]. The three tag elements are: (1) ENAMEX used to represent persons' names, locations, and organizations, (2) NUMEX used to represent numerical expressions, money and percentages, and (3) TIMEX which is used to represent time and date expressions. The second tag set was initiated in 2002 during the Conference on Computational Natural Language Learning [5]. In this tag set, NEs are classified into four categories; person name, location, organization, and miscellaneous. Using this scheme, chunks of NEs in a dataset are tagged using the Inside-Outside-Beginning (IOB) format, where a token is tagged with "B" to indicate that it is at the beginning of a chunk, with "I" to indicate that it is inside the chunk and with "O" to indicates that the token does not belong to a chunk (i.e. outside, not part of the chunk). The third tag set was initiated in 2003 by the Automatic Content Extraction (ACE) program [6]. The tag set classifies NEs into four categories: person name, facility, organization, and geographical and political entities (GPE). Vehicles and weapons were added to the tag set as two new categories in ACE 2004 and ACE 2005.

NER approaches can be categorized into hand-made rule-based NER, machine learning (ML) NER and hybrid NER [7]. In the hand-made rule-based approach, NER is performed using human devised rule sets, while in ML approach, the NER problem is converted into a classification problem and hence ML techniques are used as a solution. In the hybrid NER approach, a combination of rule-based and ML-based approaches are used together employing the best of each. This paper proposes a ML based technique that exploits multilingual Wikipedia for the purpose of building Arabic NE gazetteer lists that will help improve Arabic NER. The rest of this paper is structured as follows: section 2 discusses related works, section 3 describes the methodology, section 4 presents the

experimental results and evaluation, and section 5 is the conclusion.

II. Related Work

In this section, a description of the different Arabic NER techniques that have been proposed is provided. One straightforward NER technique was proposed by Shah et al. in [7], who devised SYNERGY, an Arabic NER system that translates Arabic to English before performing NER. Other techniques make use of parallel corpora. Samy et al. in [9] used parallel corpora in Spanish and Arabic, and a Spanish NE tagger to tag the name entities in the Arabic corpus. In their approach, Spanish NEs are extracted from Spanish sentences and classified into sub-lists according to their type. Date NEs are passed to the date module, while other types such as person, location, geographical names and some acronyms are passed to the transliteration module. Although the authors reported high precision and recall, it should be noted that their approach is applicable only when a parallel corpus is available. Darwish and Gao in [10] proposed multiple approaches to improve NER from microblogs. This approach is language independent and comprises of creating large gazetteers, domain adaption and a two-pass semi-supervised method.

A. Ruled-based Approaches

The work of Mesfar in [11] is an example of the Arabic rule-based approach. The system the researcher described, combines a morphological parser and a syntactic parser built with the NooJ linguistic development environment. The system starts by tokenizing text, then this text is sent to the morphological analyzer which uses finite state technology to parse vowelized, partially vowelized and un-vowelized text. The recognized forms associated with linguistic information are sent to an Arabic NER system, which in turn recognizes the NE's with the help of knowledge sources such as gazetteers and grammars. Mesfar's system uses the ENAMEX, TIMEX and NUMEX tagging scheme.

Shaalan and Raza in [12] developed PERA, a rule-based person NER system for Arabic language. It consists of a lexicon in the form of gazetteer name lists, and a grammar in the form of regular expressions. The authors improved their work by proposing a modified system called NERA, a rule-based Name Entity Recognition for Arabic, consisting of a dictionary of names (whitelist) and grammar in the form of regular expressions. The system is capable of recognizing and extracting person name, location, company, date, time, price, measurement, phone number, ISBN and file name [13], [14]. Shihadeh and Neumann in [15] developed another Arabic NER system named ARNE. Their work performs tokenization, morphological analysis, Buckwalter transliteration, POS tagging and finally the recognition of NEs is performed using the Inside-Outside-Beginning tagging scheme.

Zaghouani in [16] proposed RENAR, a rule-based Arabic NE recognition system. RENAR uses a freely available corpus and other resources that are built by the author, such as stop words list, modifiers lists, and person, location, organization gazetteers. This system comprises of three main steps; the pre-processing, lookup of known names and finally the local grammar step which is responsible of recognizing the unknown names. This system is a multilingual NER system used to extract three Arabic NEs; person, location and organization.

An Arabic NER method based on transducer cascade is proposed by Mesmia et al. in [17]. Their method consists of three main steps: firstly, the construction of two dictionaries that contain the first names and the last names. Secondly, the identification of extraction rules. And thirdly, the establishment of transducers. The testing of this system was done using a Wikipedia corpus, which is constructed using the Arabic kiwix tool.

Two other research efforts employed rule based approaches for domain specific Arabic NER; one of which targeted the crime domain and was proposed by Asharef et al. in [18], while the other targeted the political domain and was proposed by Alshref and Aziz in [19].

B. Machine Learning-Based Approaches

Many Arabic NER research papers fall under the ML category. Mohammed and Omar in [20] proposed an Arabic NER system based on Artificial Neural Networks (ANN) that aims to classify Arabic NEs. Their system consists of three stages. In the first stage, the text is pre-processed in order to clean the collected data. In the second stage, Arabic letters are converted to the Roman alphabet. Finally, in the third stage the data is classified using ANNs. The accuracy of their system reached 92%. This result was compared with the result obtained by the decision trees (DTs) which reached 87% when applied on the same data.

A semi-supervised algorithm for Arabic NER known as ASemiNER was proposed by Althobaiti et al. in [21]. ASemiNER, does not require annotated training data or gazetteers and can recognize three NEs; person, location and organization. This algorithm consists of three main components that attempt to extract semantic information from natural text; first the pattern induction and consists of initial patterns and generalization steps. Second, instance extraction. And third, instance ranking/selection. Another approach that combines the semi-supervised and the distant learning techniques was proposed by Althobaiti et al. in [22]. This technique is capable of recognizing three NEs; person, location and organization. The two classifiers, semi-supervised and the distant learning, are trained and combined using the Bayesian Classifier Combination (BCC) procedure.

NAMERAMA is another system that recognizes Arabic NEs in the medical domain [23]. It is based on Bayesian Belief Network (BBN) and uses the Inside-Outside tagging scheme to

identify disease names, symptoms, treatment methods, and diagnosis methods. NAMERAMA comprises four steps; pre-processing, data analysis, feature extraction and classification.

Benajiba et al. in [24] presented another system based on maximum entropy (ME). The authors developed their own corpus known as ANERcorp and their own gazetteers known as ANERgazet. A two-step improvement to this system is proposed by Benajiba and Rosso in [25]; the first step concentrates on the delimitation of the NE's using the contextual and POS-tag information, while the second step is fully ME-based. A further enhancement on the accuracy of ANERsys was presented by using Conditional Random Fields instead of the Maximum Entropy probabilistic model [26].

A NER system which uses support vector machine (SVM), together with language independent and language dependent features was described by [27]. The system uses the Inside-Outside-Beginning tagging scheme. An approach based on SVM is proposed by [28], in order to recognize person, location and organization named entities. The approach combines publicly available systems and corpora; such as YamCha tool, Buckwalter Arabic Morphological Analyzer (BAMA), the Stanford POS tagger, and ANERgazet. Another SVM based approach namely ANER is proposed in [29], which uses Hidden Markov Model (HMM) and a combination of binary features, in addition to the Inside-Outside-Beginning tagging scheme.

C. *Hybrid Approaches*

Only few research efforts in literature detailed hybrid Arabic NER approaches. Shaalan and Oudah in [30] proposed a system that integrates ML with rule-based approaches. The system consists of three main phases; (1) a rule-based NER phase, (2) a feature selection and extraction phase, and a (3) ML phase. The authors identified 11 types of Arabic name entities: Person, Location, Organization, Date, Time, Price, Measurement, Percent, Phone Number, ISBN and File Name.

Abdallah et al. in [31] proposed a hybrid Arabic NER system that combines NERA with DTs. Their system works sequentially by using the results of the rule-based system NERA as an input features for the ML classifiers, the DTs. This system focused on three NEs; person, location and organization. Another hybrid system that combined rule-based with SVM was proposed by Meselhi et al. in [32] to recognize eight NEs; Location, Person, Organization, Date, Time, Price, Measurement and Percent. The components of Meselhi's system work in parallel.

III. METHODOLOGY

This section describes the different steps undertaken to build the Arabic NE gazetteer lists. These steps are organized in three consecutive phases: (1) corpus building phase, (2) testing dataset building phase, and (3) building the gazette. The individual steps undertaken in each phase are described below.

A. *Phase 1: Corpus building*

This phase consists of three main steps:

Step 1 Collecting data: 300 NEs were carefully selected manually from Wikipedia. These NEs represent titles of 300 Wikipedia pages in Arabic language. Since this research concerns three categories, the 300 NEs were distributed equally amongst the three categories; person, location, and organization. Thus, each category is made of 100 NEs. Table IV in appendix A shows a list of the selected NEs.

Step 2: Fetching pages from Wikipedia: For each NE collected in step 1, the Wikipedia pages were fetched and only the textual content of these pages was taken into consideration, ignoring links, figures, and tables. Similarly, each group of 100 files was annotated with a specific NE class.

Step 3: Processing pages: Prior to training, each file content was processed by eliminating punctuation, stop words, non-Arabic text, and diacritical marks. Finally, stemming was performed.

B. *Phase 2: Dataset preparation*

The dataset used in this work is created and prepared for testing purposes. This phase consists of five main steps:

Step 1: Collecting data: Text is collected from Aljazeera in English language. The collected text includes approximately 100 NEs.

Step 2: Extracting English Nes: Stanford NER is applied on the English text to extract English NEs. Each NE is assigned one of the NE tags corresponding to three classes; person, location, and organization. For example, Jordan is assigned the tag location.

Step 3: Fetching English Wikipedia titles: For each English NE extracted in the previous step, a Wikipedia search is performed using the extracted NE as the search term. For example, Jordan is looked up in Wikipedia, if there exists a page entitled Jordan in Wikipedia, then the page is fetched.

Step 4: Checking for Arabic parallel page: From the Wikipedia page of each English title obtained in the previous step, an existence check of its Arabic parallel page is performed. If the page exits, then the Arabic Wikipedia title and page are fetched. Otherwise, the language of the Wikipedia is changed to Arabic and the English title is used to search for Arabic titles and pages. Then the textual content of these pages are grouped according to the NE class given by the Stanford NER into three classes. For example: check if the page Jordan contains an Arabic language link, if yes then fetch the Arabic page for it, otherwise search for Jordan using Arabic Wikipedia

interface, in either way, select the textual content of the Arabic parallel page and save it with its original tag.

Step 5: Processing pages: Each Arabic file content is then processed and prepared for testing. The processing consists of similarly punctuation, stop words, non-Arabic text, and diacritical marks removal, followed by stemming. It is important to note that the Stanford NER recognizes only single tokens and not chunks of NEs, while a Wikipedia search for these tokens fetches chunks of NEs, thus the original token may not be equivalent to the title fetched. For example, Sumaya represents a person and is assigned the NER tag: person. The proposed system may fetch the Arabic parallel Wikipedia page: Princess Sumaya University for Technology which represents a location. Thus, phase 3 is applied to verify the NER tag change.

C. *Phase 3: Resource building*

This phase is the final phase and which involves the gazette building. In the gazette building phase, a naïve Bayes (NB) classifier is used to assign NE tags to the Arabic NE. NB classifier is considered one of the probabilistic classifiers. NB classifier is based on Bayes theorem shown below and independent assumptions between features.

$$\Pi(A|B) = (\Pi(B|A)\Pi(A))/\Pi(B) \quad (1)$$

Where A and B are events, $\Pi(B) \neq 0$, $\Pi(A)$ and $\Pi(B)$ are the probabilities of observing A and B. $\Pi(A|B)$ is a conditional probability, which means the probability of A given that B is true. $\Pi(B|A)$ is the probability of event B given that A is true. Three main steps are conducted: corpus training, testing data training, and finally the classifying step in which the Arabic title is assigned either the same NER tag given by the Stanford NER or a new one.

Step 1: Corpus Training; The content of each processed file for each category (i.e. person, location, and organization) in the corpus building phase is converted into tuples of (word, label), in which the label is the original tag assigned to the document. Next the features are extracted from these contents. As previously mentioned in the methodology section, it is noted that processing is done on the file contents such as removing punctuation, stop words, non-Arabic text, diacritical marks, and stemming. This processing facilitates feature extraction. Each word in the file is then given a frequency and sent to the NB classifier for training.

Step 2: Test Data Training; The same processing steps done in the previous step (i.e. step 1 of phase 3) are applied on the content of each processed file for each category (i.e. person, location, and organization) in the testing dataset. Each word in the file is then assigned a frequency.

Step 3: Classifying; The test data is sent to the NB classifier, where the document is either assigned to a new NER class or the original class is kept according to the training it is given.

IV. Experimental Results and Evaluation

This section describes the experimental results and the evaluation process which was done manually due to lack of available resources. First, when building the corpus, the 300 NEs were selected carefully by first taking into consideration their existence in Wikipedia, and second checking if the page contains a reasonable text. Table IV in appendix A represents the selected NEs for each category.

Second, the testing data is created from English selected text from Aljazeera.com. This text contains at least 100 NE, recognized and assigned a specific NER class by the Stanford NER, and then the Arabic Wikipedia pages were fetched for these English NEs. It is important to note that the disambiguation pages were neglected. Table I shows the exact number of NEs in each class in the testing dataset.

Third, the Arabic NEs are sent to the classifier, to assign it to a NER class. Table V in appendix A shows the resulted NEs for the testing and their original NER classes, new NER classes, and the correction done manually. Original class is the one given by the Stanford NER and the new class is the one given by the classifier. It is also important to note that the evaluation is done on the dataset before and after stemming, but the results were similar in both cases.

Finally, the evaluation is done manually. As shown in table V in appendix A, 7 NEs in the testing dataset had ORG as the original tag when it should have been LOC after applying the classifier, these NEs were correctly classified as LOC. Another 5 NEs in the testing dataset had PER as the original tag. Similarly, these 5 NEs after applying the NB classifier were correctly classified into LOC. Table II illustrates the accuracy, recall, and precision obtained from the manual evaluation for the proposed classifier, as noted the accuracy, recall and precision are 52.75%, 17.3% and 33.3% respectively, indicating that classifier is poor in classifying person and organization. However, in regarding to classifying LOC NEs, the accuracy reached 80.37% with 52% recall and 100% precision, this is illustrated in table III.

The poor classification of both person and organization is because the contents of their pages in Wikipedia are very diverse and might need to have more fine grain classes, or the need for a bigger corpus. However, location pages contain a uniform and frequently occurring words.

TABLE I. Arabic NEs training dataset statistics

PERSON	LOCATION	ORGANIZATION	OTHER	Total
28	48	15	16	107

TABLE II. Arabic NEs statistical measures for classifying LOC, ORG and PER

Measure	%
Accuracy	52.75%
Recall	17.3%
Precision	33.3%

TABLE III. ARABIC NES STATISTICAL MEASURES FOR CLASSIFYING LOC

Measure	%
Accuracy	80.37%
Recall	52%
Precision	100%

V. CONCLUSION

This paper focused on building Arabic named entity lexicon, by using the English name entities to exploit Wikipedia for Arabic name entities, and classifying them into three main categories; person, location, and organization by applying the Naïve Bayes classifier. The process starts with building two corpora's, one for training and the other for testing. The evaluation is done manually because Arabic language lacks for such resources and the accuracy, recall and precision values obtained are 52.75%, 17.3% and 33.3% respectively, this low number is caused by the poor classification of the classes person and organization, while classifying the class location alone have the accuracy reached 80.37%.

APPENDIX A

TABLE IV. ARABIC NES CONTAINED IN THE CORPUS

No.	PERSON	LOCATION	ORGANIZATION
1	الأميرةسميةبنتالحسن	جبلالشعانبي	فيسبوك
2	وداداللكيلاني	المحيطالهادئ	تويتر
3	ناميهامورو	صحراءخلوص	إنستغرام
4	شهدةبنتأحمدالإبريالدينوري	محميةرأسمحمد	بدرالدينللبترول (بابتيكو)
5	أمسلمة	إسبانيا	إعمار (شركة)
6	سابرينافيريللي	سوريا	مجموعةتزين
7	سابينليزيكي	الاردن	شركةالنفطالوطنيةالعراقية
8	ابتسامالطفي	كابول	شركةكيانالسعوديةللبتروكيماويات
9	ابتسامهجرس	ألبانيا	شركةعامة
10	ابتهاجمحمد	الجزائر	قطرللبترول
11	غادةالسمان	ساموا الأمريكية	شركةبترولأبوظبيالوطنية
12	ريااأبيراشد	أندورا	الشركةالوطنيةللبتروكيماويات
13	ثرياأغاأوغلو	أنغولا	تسو (تواصلاجتماعي)
14	ثرياالشاوي	أنغويلا	خدمةالشبكةالاجتماعية
15	جميلةبوحيرد	أنتاركتيكا	مايسبيس
16	سارة	أنتيغواوبربودا	أليكساإنترنت
17	آسياداغر	الأرجنتين	أمازون (شركة)
18	آلاءمرابط	أرمينيا	سنابشات
19	مروةمحمد	أروبا	فكونتاكتي
20	سميرةسعيد	أستراليا	أسكإفام
21	منتهىالرمحي	النمسا	جودريدز
22	منتهىمحمدرحيم	أذربيجان	فريندستير
23	سلوىبنتعبداللهالهزاع	الباهاماس	إيباي
24	سلوىالجسار	البحرين	ياهو!
25	دريةشرفالدين	بنغلاديش	سينا (شركة)
26	سهيرالعلي	بربادوس	مايكروسوفت
27	معصومةالمبارك	روسياالبيضاء	لينكدإن
28	آمالكربول	بلجيكا	بحثجوجل
29	حكيمةالحيطي	بليز	تاوباو (موقع)
30	رانياالعبدالله	بنين	ويكيبيديا
31	هيابنتالحسين	جزربرمود	يوتيوب
32	دياناكرزون	بوتان	إمإسإن
33	رانياالكردي	بوليفيا	أبل
34	ميسحمدان	البوسنةوالهرسك	بايبال
35	علاالفارس	بوتسوانا	عليإكسبريس (موقع)
36	حكيمابوالقاسمفردوسيبطوسي	البرازيل	الجيشالسوري
37	طارقالبشري	بروناي	ناديشبابالتغيير
38	غلينجونسون	بلغاريا	حزبمصرالثورة
39	غايتانوشيريا	بوركينافاسو	الحزبالديمقراطيالكردستاني
40	تشاكبولانيك	بوروندي	الحزبالشيوعيالعراقي
41	أبومنصورمحمدالقاهربالله	كمبوديا	الأممالمتحدة
42	عبدالفتاحالسيسي	كاميرون	نمورالتاميل
43	سيديمحمدولدالشيخ عبدالله	كندا	جماعةأبوسياف
44	جابررزقالفولي	الرأسالأخضر	جبهةالنصرة
45	صالحالعلي	جمهوريةأفريقياالوسطى;	الجيشالسوريالحر
46	إسماعيل	تشاد	تنظيمالقاعدة
47	تحسينشردم	تشيلي	حزبالله
48	طلعتعفيفي	جمهوريةالصينالشعبية	حركةحماس
49	عبداللهبنجحش	كولومبيا	الحشدالشعبي
50	محمدمصطفهدارة	جزرالقمر	فيلقالقدس
51	محمدمرسي	جمهوريةالكونغوالديمقراطية	الحلفالأطلسي
52	شوقيضيف	جمهوريةالكونغو	جبهةالعملالإسلامي (الأردن)
53	علاءعبدالفتاح	جزركوك	الإخوانالمسلمون
54	المغيرةبنشعبة	كوستاريكا	حزبالحريةوالعدالة
55	أبوالمنصورالفضلالمسترشدبالله	ساحلالعاج	درعالجزيرة
56	إبراهيمالسلقيني	كرواتيا	تنظيمداعش - ولايةسيناء
57	إبراهيمالأول	كوبا	حوثيون
58	طارقالعلي	قبرص	تنظيمالدولةالإسلامية (داعش)
59	أبوإسلاممأحمدعبدالله	الجمهوريةالتشيكية	الجيشالسوريالحر
60	رياضنيقولا	تيرانا	حزبالله
61	جمالالدجاني	الجزائرالعاصمة	حركةحماس
62	عبداللطيفالبغدادي	اندورالافيلا	الدستور (جريدةأردنية)
63	عبدالحكيمعامر	لواندا	العربالليوم (جريدة)
64	حسنإبراهيم (ضابط)	سانجونز	الدستور (جريدةمصرية)
65	عليصبري (سياسي)	بوينسأيرس	الدستور (جريدةعراقية)
66	محمدأنورالسادات	يريفان	صحيفةالشرق (السعودية)
67	عدليمنصور	أورنجستاد	صحيفةالشرق (قطر)
68	جمالعبدالناصر	كانبرا	الوطن (جريدةبحرينية)
69	صدامحسين	فيينا	الوطن (جريدةتونسية)
70	شيجينبينغ	باكو	الوطن (جريدةجزائرية)
71	هوجينتاو	ناساو	الجمعيةالثقافيةالعلميةجنو/لينوكس والبرمجياتالحرة
72	كريمماسيموف	المنامة	الجمعيةالثقافيةالسريانيةفيسوريا
73	عبدالحميدالبكوش	دكا	الجمعيةالثقافيةبالعمار
74	معمرالقذافي	بريدجتاون	جمعيةمصرللثقافةوالحوار
75	عبداللهالثانيبنالحسين	مينسك	الجمعيةالثقافيةالاجتماعيةالنسائية

	ين		
76	شيخ	بروكسل	الجمعيةالعربيةالسعوديةللثقافةوالف نون
77	ملا	بلموبان	ندوةحريةالصحافةفيليبيا
78	أمير	بورتونوفو	مهرجانالدوخلة
79	أميرالمؤمنين	هاميلتون	معرضالرياضالدوليللكتاب
80	عاهل	ثيمفو	تكيةأمعلي
81	رتبةعسكرية	لاباز	جمعيةصناعالحياةبالأردن
82	سياسي	سراييفو	مؤسسةبيلوماليندا غيتس
83	لاعبوسط (كرةقدم)	جابورون	هيومنرايتسووتش
84	تاجر	برازيليا	الهلالالأحمرالسعودي
85	مدرس	بندرسريبجاوان	الجمعيةالخيريةالشركسيةفيالأرد ن
86	نابغةشطرنج	صوفيا	الجمعيةالخيرية (الكويت)
87	معلقرياضي	واجادوجو	جمعيةالمقاصدالخيريةالإسلامية
88	رياضي	بوجومبورا	صناعالحياة
89	مذيع	بنومبنه	الهيئةالخيريةالإسلاميةالعالمية
90	مذيعأخبار	ياوندى	جمعيةالصليبوالهلالالأحمر
91	رئيس	أوتاوا	اللجنةالدوليةللصليبالأحمر
92	الرئيسالتنفيذي	برايا	الاتحادالدوليلجمعياتالصليبالأحمر والهلالالأحمر
93	مديرتنفيذي	موريتانيا	منظمةخيرية
94	رئيس-مديرعام	انجمينا	جمعيةالشارقةالخيرية
95	باحث	سانتياغو	قريةالأطفالإسأوإس
96	باحثمابعدالدكتوراه	بكين	الجمعيةالسعوديةالخيريةلمرضا لإيدز
97	عالم (صفة)	بوغوتا	منظمةغيرربحية
98	رئيسالوزراء	موروني	مؤسسةمحمدالخامسللتضامن
99	حاكمالدولة	كينشاسا	جيشالخلاص
100	رئيسالجمهورية	برازافيل	جمعيةالمقاصدالخيريةالإسلامية

TABLE V. ARABIC NES RESULTED FROM ENGLISH TEXT WITH THEIR NER CLASS, WHERE ORIGINAL NER CLASS GIVEN BY THE STANFORD NER CLASSIFIER AND THE NEW NER CLASS GIVEN BY THE PROPOSED CLASSIFIER, AND THE CORRECTION I DONE MANUALLY.

No.	Arabic NE	Original NER class	New NER class	Correction
1	الاردن	LOC	LOC	-
2	روسيا	LOC	LOC	-
3	الطريق إلى الفلوجة (فيلم)	LOC	LOC	-
4	العراق	LOC	LOC	-
5	إسلام	LOC	LOC	-
6	بغداد	LOC	LOC	-
7	الموصل	LOC	LOC	-
8	مدينة	LOC	LOC	-
9	قلعة أربيل	LOC	LOC	-
10	القاهرة	LOC	LOC	-
11	مصر	LOC	LOC	-
12	قطر	LOC	LOC	-
13	مربع	LOC	LOC	Shape
14	أمستردام	LOC	LOC	-
15	هولندا	LOC	LOC	-
16	أوروبا	LOC	LOC	-
17	حلب	LOC	LOC	-
18	نايميخن	LOC	LOC	-
19	سوريا	LOC	LOC	-
20	هوفهايم (بافاريا)	LOC	LOC	-
21	دمشق	LOC	LOC	-
22	هولاند باتينت (نيويورك)	LOC	LOC	-
23	لبنان	LOC	LOC	-
24	برلين	LOC	LOC	-
25	الشيشان	LOC	LOC	-
26	مشرع العين	LOC	LOC	-
27	أحمد الباشا	LOC	LOC	PER
28	اليمن	LOC	LOC	-
29	بج بن عدن	LOC	LOC	-
30	الكويت	LOC	LOC	-
31	تويتر	LOC	LOC	ORG
32	برينس	LOC	LOC	PER
33	السعودية	LOC	LOC	-
34	اتحاد شعب الجزيرة العربية	LOC	LOC	ORG
35	جدة	LOC	LOC	-
36	إمارة دبي	LOC	LOC	-
37	تركيا	LOC	LOC	-
38	إندونيسيا	LOC	LOC	-
39	الولايات المتحدة	LOC	LOC	-
40	المملكة المتحدة	LOC	LOC	-
41	شمال	LOC	LOC	Direction
42	معركة طرف الغار	LOC	LOC	Battle
43	الإمارات العربية المتحدة	LOC	LOC	-
44	ناسا	ORG	LOC	ORG
45	فيلم	ORG	LOC	Movie
46	شمال	ORG	LOC	Direction
47	الولايات المتحدة	ORG	LOC	-
48	إسلام	ORG	LOC	Religion
49	نقش بارز	ORG	LOC	Art
50	كندا	ORG	LOC	-
51	المملكة المتحدة	ORG	LOC	-
52	فيس بوك	ORG	LOC	ORG
53	تنظيم الدولة الإسلامية (داعش)	ORG	LOC	ORG
54	وكالة حماية البيئة الأمريكية	ORG	LOC	ORG
55	لاجئ	ORG	LOC	PER
56	قائمة بلديات فلوريدا	ORG	LOC	-
57	الأمم المتحدة	ORG	LOC	ORG
58	الطريق إلى الفلوجة (فيلم)	ORG	LOC	Movie
59	ملف استنادي دولي افتراضي	ORG	LOC	ORG
60	معهد هولندا لتاريخ الفن	ORG	LOC	-
61	منظمة الشرطة الجنائية الدولية	ORG	LOC	ORG
62	اللجنة العربية العليا	ORG	LOC	ORG
63	صحفي	ORG	LOC	PER
64	لغة عربية	ORG	LOC	Language
65	مسلم	ORG	LOC	PER
66	عربة نقل	ORG	LOC	Transport
67	بنغال	ORG	LOC	-
68	وكالة الطاقة النووية	ORG	LOC	ORG
69	فريق أول (رتبة عسكرية)	ORG	LOC	PER
70	ذكاء	ORG	LOC	PER
71	رويترز	ORG	LOC	ORG
72	اليمن	ORG	LOC	-

73	بريد إلكتروني	ORG	LOC	Address
74	هيسبانو سويسا	ORG	LOC	ORG
75	تويتر	ORG	LOC	ORG
76	برينس	ORG	LOC	PER
77	غزوات المغول للشام	PER	LOC	Battle
78	شيخ احمد (غربي أردبيل)	PER	LOC	-
79	طلعت عفيفي	PER	LOC	PER
80	وداد الكيلاني	PER	LOC	PER
81	إسماعيل	PER	LOC	PER
82	إليس غروف (إلينوي)	PER	LOC	PER
83	محمد محمود عبد العزيز	PER	LOC	PER
84	عبد الفتاح السيسي	PER	LOC	PER
85	زكريا	PER	LOC	PER
86	محمد	PER	LOC	PER
87	معدة	PER	LOC	Organ
88	ملا	PER	LOC	PER
89	قوس (ترقيم)	PER	LOC	Punctuation
90	حوثيون	PER	LOC	ORG
91	سيدي محمد ولد الشيخ عبد الله	PER	LOC	PER
92	شيخ	PER	LOC	PER
93	عبد المالك الدهامشة	PER	LOC	PER
94	محمد أحمد المخلافي	PER	LOC	PER
95	طارق العلي	PER	LOC	PER
96	برج العرب (فندق)	PER	LOC	-
97	سانتياغو	PER	LOC	-
98	واد	PER	LOC	-
99	حيدر العبادي	PER	LOC	PER
100	رمضان	PER	LOC	Month
101	حسين صادق المصراتي	PER	LOC	PER
102	محمد مرسي	PER	LOC	PER
103	شوقي ضيف	PER	LOC	PER
104	ثوم (جنس)	PER	LOC	Plant
105	هلال (مقاطعة تشناران)	PER	LOC	-
106	علاء عبد الفتاح	PER	LOC	PER
107	أسماء محفوظ	PER	LOC	PER

REFERENCES

[1] Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato,and Juan Miguel Gómez-Berbís. "Named entity recognition: fallacies, challenges and opportunities." Computer Standards & Interfaces 35, no. 5 (2013): 482-489.

[2] Leser, Ulf, and Jörg Hakenberg. "What makes a gene name? Named entity recognition in the biomedical literature." Briefings in bioinformatics 6, no. 4 (2005): 357-369.

[3] Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. Comput. Linguist, 40(2):469–510, June.

[4] The 6th Message Understanding Conference (MUC-6). 1995.http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

[5] the Conference on Computational Natural Language Learning (CoNLL). 2002. http://ifarm.nl/signll/conll/

[6] The Automatic Content Extraction (ACE) tag sets for English, Arabic, and Chineseare available at http://projects.ldc.upenn.edu/ace/data/

[7] Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat. "Named entity recognition approaches." International Journal of Computer Science and Network Security 8, no. 2 (2008): 339-344.

[8] Shah R., Lin B., Gershman A. and Frederking R. (2010). SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation. In Proceedings of (AfLaT), LREC, Valleta, Malta.

[9] Samy, Doaa, Antonio Moreno, and Jose M. Guirao. "A proposal for an Arabic named entity tagger leveraging a parallel corpus." In International Conference RANLP, Borovets, Bulgaria, pp. 459-465. 2005.

[10] Darwish, Kareem, and Wei Gao. "Simple Effective Microblog Named Entity Recognition: Arabic as an Example." In LREC, pp. 2513-2517. 2014.

[11] Mesfar, Slim. "Named entity recognition for arabic using syntactic grammars." In Natural Language Processing and Information Systems, pp. 305-316. Springer Berlin Heidelberg, 2007.

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

[13] Shaalan, Khaled, and Hafsa Raza. "Arabic named entity recognition from diverse text types." In Advances in Natural Language Processing, pp. 440-451. Springer Berlin Heidelberg, 2008.

[14] Khaled Shaalan, Hafsa Raza, NERA: Named Entity Recognition for Arabic, Journal of the American Society for Information Science and Technology, v.60 n.8, p.1652-1663, August 2009.

[15] Carolin Shihadeh and Günter Neumann. ARNE - A tool for Namend Entity Recognition from Arabic Text. Inproceedings of Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4) 2012. AMTA.

[16] Zaghouani, Wajdi. "RENAR: A rule-based Arabic named entity recognition system." ACM Transactions on Asian Language Information Processing (TALIP) 11, no. 1 (2012): 2.

[17] Mesmia, Fatma Ben, Nathalie Friburger, Kais Haddar, and Denis Maurel. "Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia." RECENT ADVANCES IN: 48. 2015.

[18] Asharef, M., Nazlia Omar, M. Albared, Zhou Minhui, Wang Weiming, and Zhou Jingjing. "Arabic named entity recognition in crime documents." Journal of Theoretical and Applied Information Technology 44, no. 1 (2012): 1-6.

[19] Alshref, Mhamed, H. Halema, and Mohd Juzaiddin Ab Aziz. "Named entity recognition for political domain in Arabic language." Asian Journal of Applied Sciences 7, no. 1 (2014): 13-21.

[20] Mohammed, Naji F., and Nazlia Omar. "Arabic named entity recognition using artificial neural network." Journal of Computer Science 8, no. 8 (2012): 1285.

[21] Althobaiti, Maha, Udo Kruschwitz, and Massimo Poesio. "A Semi-supervised Learning Approach to Arabic Named Entity Recognition." In RANLP, pp. 32-40. 2013.

[22] Althobaiti, Maha, Udo Kruschwitz, and Massimo Poesio. "Combining Minimally-supervised Methods for Arabic Named Entity Recognition."Transactions of the Association for Computational Linguistics 3 (2015): 243-255.

[23] Alanazi, Saad, Bernadette Sharp, and Clare Stanier. "A Named Entity Recognition System Applied to Arabic Text in the Medical Domain."International Journal of Computer Science Issues (IJCSI) 12, no. 3 (2015): 109.

[24] Benajiba, Yassine, Paolo Rosso, and José Miguel Benedíruiz. "Anersys: An arabic named entity recognition system based on maximum entropy."

InInternational Conference on Intelligent Text Processing and Computational Linguistics, pp. 143-153. Springer Berlin Heidelberg, 2007.

[25] Benajiba, Yassine, and Paolo Rosso. "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.

[26] Benajiba, Yassine, and Paolo Rosso. "Arabic named entity recognition using conditional random fields." In Proc. of Workshop on HLT & NLP within the Arabic World, LREC, vol. 8, pp. 143-153. 2008.

[27] Benajiba, Yassine, Mona Diab, and Paolo Rosso. "Arabic named entity recognition: An svm-based approach." In Proceedings of 2008 Arab International Conference on Information Technology (ACIT), pp. 16-18. 2008.

[28] O'Steen, David, and David Breeden. "Named Entity Recognition in Arabic: A Combined Approach." BA (Hons), Stanford University, USA (2009).

[29] Koulali, Rim, and Abdelouafi Meziane. "A contribution to Arabic named entity recognition." In ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on, pp. 46-52. IEEE, 2012.

[30] Shaalan, Khaled, and Mai Oudah. "A hybrid approach to Arabic named entity recognition." Journal of Information Science 40, no. 1 (2014): 67-87.

[31] Abdallah, Sherief, Khaled Shaalan, and Muhammad Shoaib. "Integrating rule-based system with classification for arabic named entity recognition." InInternational Conference on Intelligent Text Processing and Computational Linguistics, pp. 311-322. Springer Berlin Heidelberg, 2012.

[32] Meselhi, Mohamed A., Hitham M. Abo Bakr, Ibrahim Ziedan, and Khaled Shaalan. "Hybrid Named Entity Recognition-Application to Arabic Language." In Computer Engineering & Systems (ICCES), 2014 9th International Conference on, pp. 80-85. IEEE, 2014.