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, Jordan

Abstract: *This paper focuses on the creation of Arabic named entity gazetteers, 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 datasets. The dataset for the training is constructed using only Arabic text, whereas, the testing dataset is derived from an English text using the Stanford name entity recognizer. 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.*

Received February 7, 2017; accepted May 10, 2017

1. Introduction

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 [15]. Additional classes can include biological species, genes, proteins, diseases, and anatomy [13]. 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 [22].

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 sixth Message Understanding Conference (MAC-6)[1]. 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[2]. 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[3]. 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 [14]. 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

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

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

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

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 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.

2. Related Work

In this section, a description of the different Arabic NER techniques that have been proposed is provided. One straight forward NER technique was proposed by Shah *et al*. [27], 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*. [21] 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 were extracted from Spanish sentences and classified into sub-lists according to their type. Date NEs were passed to the date module, while other types such as person, location, geographical names and some acronyms were passed to the transliteration module. Although the authors reported high precision and recall, it should be noted that their approach was applicable only when a parallel corpus is available. Darwish and Gao in [11] proposed multiple approaches to improve NER from microblogs. This approach is language independent and comprises of three main steps, firstly, creating of large gazetteers, secondly, domain adaption is applied and thirdly a two-pass semi-supervised method is applied.

2.1. Ruled-based Approaches

The work of Mesfar [17] 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 analyser which uses finite state technology to parse vowelized, partially vowelized and un-vowelized text. The recognized forms associated with linguistic information were 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 used the ENAMEX, TIMEX and NUMEX tagging scheme.

Shaalan and Raza [25] developed PERA, a rule-based person NER system for Arabic language. It consists of a lexicon in the form of name gazetteer, 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 [24, 26]. Shihadeh and Neumann [28] developed another Arabic NER system named ARNE. Their work performs tokenization, morphological analysis, Buckwalter transliteration, POS tagging and finally, the recognition of NEs was performed using the Inside-Outside-Beginning tagging scheme.

Zaghouani [29] proposed RENAR, a rule-based Arabic NE recognition system. RENAR uses a freely available corpus and other resources that were 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*. [18]. 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*. [6], while the other targeted the political domain and was proposed by Alshref and Aziz [3].

2.2. Machine Learning-Based Approaches

Many Arabic NER research papers fall under the ML category. Mohammed and Omar [19] 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 were converted to the Roman alphabet. Finally, in the third stage the data was 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*. [4]. 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.* [5]. This technique is capable of recognizing three NEs; person, location and organization. The two classifiers, semi-supervised and the distant learning were trained and combined using the Bayesian Classifier Combination (BCC) procedure.

NAMERAMA is another system that recognizes Arabic NEs in the medical domain [1]. 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.* [10] 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 [7]; 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 [8].

A NER system which uses Support Vector Machine (SVM), together with language independent and language dependent features was described by Benajiba *et al.* [9]. The system uses the Inside-Outside-Beginning tagging scheme. An approach based on SVM is proposed by O'Steen and Breeden [20], 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 by Koulali and Meziane [12], which uses Hidden Markov Model (HMM) and a combination of binary features, in addition to the Inside-Outside-Beginning tagging scheme.

2.3. Hybrid Approaches

Only few research efforts in literature detailed hybrid Arabic NER approaches. Shaalan and Raza [24] 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.
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.* [2] 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.* [16] to recognize eight NEs; Location, Person, Organization, Date, Time, Price, Measurement and Percent. The components of Meselhi's system work in parallel.

3. Methodology

This section describes the different steps undertaken to build the Arabic NE gazetteer. These steps are organized in three consecutive phases:

1. Training dataset preparation phase.
2. Testing dataset preparation phase.
3. Building the gazette.

The individual steps undertaken in each phase are described below.

3.1. Phase 1: Training Dataset Preparation

This phase prepares a dataset to be trained and used in phase 3, and 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 4 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.

3.2. Phase 2: Testing Dataset Preparation

The testing dataset is created and prepared using five main steps:

- *Step 1. Collecting Data*: text is collected from Aljazeera website 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 an organization. Thus, phase 3 is applied to verify the NER tag change.

3.3. 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, and is based on Bayes theorem shown below and it is also based on independent assumptions between features.

$$\Pi(A|B) = (\Pi(B|A)\Pi(A))/\Pi(B) \tag{1}$$

Where A and B are events, $\Pi(A)$ 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. And $\Pi(B|A)$ is the probability of event B given that A is true. Three main steps are conducted: building the training dataset, building the testing dataset, 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.* Building the Training Dataset: the content of each processed file for each category (i.e., person, location, and organization) in the training dataset preparation 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.* Building the Testing Dataset: 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.

4. Experimental Results and Evaluation

This section describes the experiment and discusses the evaluation process, which done manually due to lack of available resources, and the experimental results. 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 4 in appendix A represents the selected NEs for each category.

Second, the testing dataset 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 1 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 5 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 5 in appendix A, seven 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 2 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 3.

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 uniform and frequently occurring words.

Table 1. Arabic NEs training dataset statistics.

PER	LOC	ORG	Other	Total
28	48	15	16	107

Table 2. Arabic NEs statistical measures for classifying LOC, ORG. and PER

Measure	%
Accuracy	52.75%
Recall	17.3%
Precision	33.3%

Table 3. Arabic NEs statistical measures for classifying LOC.

Measure	%
Accuracy	80.37%
Recall	52%
Precision	100%

5. Conclusions

This paper focused on building Arabic named entity gazetteer, 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 collecting datasets for training and 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%.

References

[1] Alanazi S., Bernadette S., and Clare S., “A Named Entity Recognition System Applied to Arabic Text in the Medical Domain,” *International Journal of Computer Science Issues*, vol. 12, no. 3, pp. 109-117, 2015.

[2] Abdallah S., Shaalan K., and Shoaib M., “Integrating Rule-Based System with Classification for Arabic Named Entity Recognition,” *in Proceeding of International Conference on Intelligent Text Processing and Computational Linguistics*, Berlin, pp. 311-322, 2012.

[3] Alshref M. and Aziz M., “Named Entity Recognition for Political Domain in Arabic Language,” *Asian Journal of Applied Sciences*, vol. 7, no. 1, pp. 13-21, 2014.

[4] Althobaiti M., Kruschwitz U., and Poesio M., “A Semi-Supervised Learning Approach to Arabic Named Entity Recognition,” *in Proceeding of International Conference on Recent Advances in Natural Language Processing*, Hissar, pp. 32-40, 2013.

[5] Althobaiti M., Kruschwitz U., and Poesio M., “Combining Minimally-supervised Methods for Arabic Named Entity Recognition,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 243-255, 2015.

[6] Asharef M., Omar N., and Albared M., “Arabic Named Entity Recognition in Crime Documents,” *Journal of Theoretical and Applied Information Technology*, vol. 44, no. 1, pp. 1-6. 2012.

[7] Benajiba Y. and Rosso P., “ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information,” *in Proceeding of 3rd Indian International Conference on Artificial Intelligence*, Pune, pp. 1814-1823. 2007.

[8] Benajiba Y. and Rosso P., “Arabic Named Entity Recognition Using Conditional Random Fields,” *in Proceeding of Workshop on HLT and NLP within the Arabic World*, pp. 143-153. 2008.

[9] Benajiba Y., Diab M., and Rosso P., “Arabic Named Entity Recognition: An SVM-Based Approach,” *in Proceeding of Arab International Conference on Information Technology*, Hammamet, pp. 16-18. 2008.

[10] Benajiba Y., Rosso P., and Benedíruiz J., “Anersys: An Arabic Named Entity Recognition System Based on Maximum Entropy,” *in Proceeding of International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico, pp. 143-153, 2007.

[11] Darwish K. and Gao W., “Simple Effective Microblog Named Entity Recognition: Arabic as an Example,” *in Preceding of 9th International*

Conference on Language Resources and Evaluation, Reykjavik, pp. 2513-2517. 2014.

[12] Koulali R. and Meziane A., "A Contribution to Arabic Named Entity Recognition," *in Proceeding of 10th International Conference on ICT and Knowledge Engineering*, Bangkok, pp. 46-52, 2012.

[13] Leser U. and Jörg H., "What Makes A Gene Name? Named Entity Recognition in the Biomedical Literature," *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 357-369, 2005.

[14] Mansouri A., Affendey L., and Mamat A., "Named Entity Recognition Approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339-344, 2008.

[15] Marrero M., Urbano J., Sánchez-Cuadrado S., Morato J., and Gómez-Berbís J., "Named Entity Recognition: Fallacies, Challenges and Opportunities," *Computer Standards and Interfaces*, vol. 35, no. 5, pp. 482-489, 2013

[16] Meselhi M., Abo Bakr H., Ziedan I., and Shaalan K., "Hybrid Named Entity Recognition-Application to Arabic Language," *in Proceeding of 9th International Conference on Computer Engineering and Systems*, Cairo, pp. 80-85, 2014.

[17] Mesfar S., "Named Entity Recognition for Arabic Using Syntactic Grammars," *in Proceeding of 12th International Conference on Applications of Natural Language to Information Systems*, Paris, pp. 305-316, 2007.

[18] Mesmia F., Haddar K., Friburger N., and Maurel D., "Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia," *in Proceeding of Recent Advances in Natural Language Processing*, Hissar, pp. 48-54. 2015.

[19] Mohammed N. and Omar N., "Arabic Named Entity Recognition Using Artificial Neural Network," *Journal of Computer Science*, vol. 8, no. 8, pp. 1285, 2012.

[20] O'Steen D. and Breeden D., "Named Entity Recognition in Arabic: A Combined Approach," Report, Stanford University, 2009.

[21] Samy D., Moreno A., and Guirao J., "A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus," *in Proceeding of International Conference on Recent Advances in Natural Language Processing*, Borovets, pp. 459-465. 2005.

[22] Shaalan K., "A Survey of Arabic Named Entity Recognition and Classification," *Computational Linguistics*, vol. 40, no. 2, pp. 469-510, 2014.

[23] Shaalan K. and Oudah M., "A Hybrid Approach to Arabic Named Entity Recognition," *Journal of Information Science*, vol. 40, no. 1, pp. 67-87, 2014.

[24] Shaalan K. and Raza H., "NERA: Named Entity Recognition for Arabic," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 8, pp.1652-1663, 2009.

[25] Shaalan K. and Raza H., "Person Name Entity Recognition for Arabic," *in Proceeding of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, pp. 17-24, 2007.

[26] Shaalan K. and Raza H., "Arabic Named Entity Recognition from Diverse Text Types," *in Advances in Natural Language Processing*, pp. 440-451, 2008.

[27] Shah R., Lin B., Gershman A., and Frederking R., "SYNERGY: a Named Entity Recognition System for Resource-Scarce Languages such as Swahili Using Online Machine Translation," *in Proceeding of the 2nd Workshop on African Language Technology*, Valleta, pp. 21-26. 2010.

[28] Shihadeh C. and Neumann G., "ARNE: A Tool for Named Entity Recognition from Arabic Text," *in Proceeding of 4th Workshop on Computational Approaches to Arabic Script-based Languages*, San Diego, pp. 24-31, 2012.

[29] Zaghouani W., "RENAR: A rule-based Arabic Named Entity Recognition System," *ACM Transactions on Asian Language Information Processing*, vol. 11, no. 1, pp. 1-13, 2012.

Mariam Biltawi is a PhD candidate in Computer Science (CS) at Princess Sumaya University for Technology (PSUT), Jordan. She received her BSc degree in CS from PSUT and her MSc degree from Balqa Applied University (BAU), Jordan, in 2005 and 2011, respectively. She worked as a computer lab supervisor in the CS department at the PSUT from November 2006 until September 2015. During her position as a lab supervisor, she worked as a part-time lecturer at the same university. Her research interests include: natural language processing, image processing, and operating systems.

Arafat Awajan is a professor of computer science at Princess Sumaya University for Technology (PSUT). He received his PhD degree in computer science from the University of Franche-Comte, France in 1987. He held different academic positions at the Royal Scientific Society and Princess Sumaya University for Technology. He was appointed as the chair of the Computer Science Department, and the chair of the Computer Graphics and Animation Department at PSUT. He had been the dean of the King Hussein School for Information Technology from 2004 to 2007, the Dean of Student Affairs from 2011- 2014 and the director of the Information Technology Center in the Royal Scientific Society from 2008-2010. He is currently the dean of the King Hussein School for computing Sciences at PSUT. His research interests include: Natural Language Processing, Arabic Text Mining and Digital Image Processing.

Sara Tedmori, In 2001, Dr. Tedmori received her BSc degree in Computer Science from the American University of Beirut, Lebanon. In 2003, she obtained her MSc degree in Multimedia and Internet Computing from Loughborough University. In 2008, she received her Engineering Doctorate in Computer Science from Loughborough University, UK. Currently she is an associate professor in the Computer Science Department at Princess Sumaya University of Technology, Jordan. Her research interests include: sentiment analysis, image processing, knowledge extraction, classification, knowledge sharing, privacy, and software engineering. She has also been involved in a number of international projects, funded mainly by the European Commission.

Akram Alkouz is an assistant professor at Princess Sumaya University for Technology. Previously, he worked as a software engineer for many tech software companies. Alkouz graduated from Technical University of Berlin with a Ph.D. degree in Computer Science. He has research interest in the fields of Data Science, Social Networks Analysis, Natural Language Processing, Machine Learning, and Information Retrieval. He has many publications in international conferences and journals.

Appendix A

Table 4. Arabic NEs contained in the corpus.

No.	Person	Location	Organization
1	الأميرة سمية بنت الحسن	جبل الشعانبي	فيسبوك
2	وداد الكيلاني	المحيط الهادئ	تويتر
3	ناميه امورو	صحراء خلوص	إنستغرام
4	شهدة بنت أحمد الإبري الدينوري	محمية رأس محمد	بدرالدين للبترول (بابتيكو)
5	أم سلمة	إسبانيا	إعمار (شركة)
6	سابرينا فيريللي	سوريا	مجموعة زين
7	سابين ليزيكي	الاردن	شركةالنفط الوطنية العراقية
8	ابتسام لطفي	كابول	شركة كيان السعودية للبتروكيماويات
9	ابتسام هجرس	ألبانيا	شركة عامة
10	ابتهاج محمد	الجزائر	قطر للبترول
11	غادة السمان	ساموا‎الأمريكية	شركة بترول أبوظبي الوطنية
12	ريا أبي راشد	أندورا	الشركة الوطنية للبتروكيماويات
13	ثريا أغا أوغلو	أنغولا	تسو (تواصل اجتماعي)
14	ثريا الشاوي	أنغويلا	خدمة الشبكة الاجتماعية
15	جميلة بوحيرد	أنتاركتيكا	مايسبيس
16	سارة	أنتيغواوبربودا	أليكسا إنترنت
17	آسيا داغر	الأرجنتين	أمازون (شركة)
18	آلاء مرابط	أرمينيا	سناب شات
19	مروة محمد	أروبا	فكونتاكتي
20	سميرة سعيد	أستراليا	آسكإفإم
21	منتهى الرمحي	النمسا	جودريدز
22	منتهى محمد رحيم	أذربيجان	فريندستير
23	سلوى بنت عبدالله الهزاع	الباهاماس	إيباي
24	سلوى الجسار	البحرين	ياهو
25	درية شرف الدين	بنغلاديش	سينا (شركة)
26	سهيرالعلي	بربادوس	مايكروسوفت
27	معصومة المبارك	روسياالبيضاء	لينكدإن
28	آمال كربول	بلجيكا	بحث جوجل
29	حكيمة الحيطي	بليز	تاوباو (موقع)
30	رانيا العبدالله	بنين	ويكيبيديا
31	هيا بنت الحسين	جزربرمود	يوتيوب
32	ديانا كرزون	بوتان	إماسإن
33	رانيا الكردي	بوليفيا	أبل
34	ميس حمدان	البوسنةوالهرسك	باييال
35	علا الفارس	بوتسوانا	علي إكسبريس (موقع)
36	حكيم ابوالقاسم فردوسى طوسى	البرازيل	الجيش السوري
37	طارق البشري	بروناي	نادي شباب التغيير
38	غلين جونسون	بلغاريا	حزب مصرالثورة
39	غايتانو شيريا	بوركينافاسو	الحزب الديمقراطي الكردستاني
40	تشاك بولانيك	بوروندي	الحزب الشيوعي العراقي
41	أبومنصور محمد القاهر بالله	كمبوديا	الأمم المتحدة
42	عبدالفتاح السيسي	كاميرون	نمورالتاميل
43	سيدي محمد ولد الشيخ عبدالله	كندا	جماعة أبوسياف
44	جابر رزق الفولي	الرأسالأخضر	جبهة النصرة
45	صالح العلي	جمهورية أفريقيا الوسطى	الجيش السوري الحر
46	إسماعيل	تشاد	تنظيم القاعدة
47	تحسين شردم	تشيلي	حزب الله
48	طلعت عفيفي	جمهورية الصين الشعبية	حركة حماس
49	عبدالله بن جحش	كولومبيا	الحشد الشعبي
50	محمد مصطفى هدارة	جزرالقمر	فيلق القدس
51	محمد مرسي	جمهورية الكونغو الديمقراطية	الحلف الأطلسي
52	شوقي ضيف	جمهورية الكونغو	جبهةالعمل الإسلامي (الأردن)
53	علاء عبدالفتاح	جزركوك	الإخوان المسلمون
54	المغيرة بن شعبة	كوستاريكا	حزب الحرية والعدالة
55	أبوالمنصورالفضل المسترشد بالله	ساحلالعاج	درع الجزيرة
56	إبراهيم السلقيني	كرواتيا	تنظيم داعش – ولاية سيناء
57	إبراهيم الأول	كوبا	حوثيون
58	طارق العلي	قبرص	تنظيم الدولة الإسلامية (داعش)
59	أبو إسلام أحمد عبدالله	الجمهورية التشيكية	الجيش السوري الحر
60	رياض نيقولا	تيرانا	حزب الله
61	جمال الدجاني	الجزائر العاصمة	حركة حماس
62	عبداللطيف البغدادي	اندورالافيلا	الدستور (جريدة أردنية)
63	عبدالحكيم عامر	لواندا	العرب اليوم (جريدة)
64	حسن إبراهيم (ضابط)	سانجونز	الدستور (جريدة مصرية)
65	علي صبري (سياسي)	بوينسآيرس	الدستور (جريدة عراقية)
66	محمد أنورالسادات	يريفان	صحيفة الشرق (السعودية)
67	عدلي منصور	أورنجستاد	صحيفةالشرق (قطر)
68	جمال عبدالناصر	كانبرا	الوطن (جريدة بحرينية)
69	صدام حسين	فيينا	الوطن (جريدة تونسية)
70	شي جينبينغ	باكو	الوطن (جريدة جزائرية)
71	هو جينتاو	ناساو	الجمعية الثقافية العلمية جنو/لينوكسوالبرمجيات الحرة
72	كريم ماسيموف	المنامة	الجمعية الثقافية السريانية فيسوريا
73	عبدالحميد البكوش	دكا	الجمعية الثقافية بالعمار
74	معمرالقذافي	بريدجتاون	جمعية مصر للثقافة والحوار
75	عبدالله الثاني بن الحسين	مينسك	الجمعية الثقافية الاجتماعية النسائية
76	شيخ	بروكسل	الجمعية العربية السعودية للثقافة والفنون
77	ملا	بلموبان	ندوة حرية الصحافة فيليبيا
78	أمير	بورتونوفو	مهرجان الدوخلة
79	أميرالمؤمنين	هاميلتون	معرض الرياض الدولي للكتاب
80	عاهل	ثيمفو	تكية أم علي

81	رتبة عسكرية	لاباز	جمعية صناع الحياة بالأردن
82	سياسي	سرايیفو	مؤسسة بيل ومليندا غيتس
83	لاعب وسط (كرةقدم)	جابورون	هيومن رايتسووتش
84	تاجر	برازيليا	الهلال الأحمرالسعودي
85	مدرس	بندرسريبيجاوان	الجمعية الخيرية الشركسية في الأردن
86	نابغة شطرنج	صوفيا	الجمعية الخيرية (الكويت)
87	معلق رياضي	واجادوجو	جمعيةالمقاصدالخيريةالإسلامية
88	رياضي	بوجومبورا	صناعالحياة
89	مذيع	بنومبنـه	الهيئة الخيرية الإسلامية العالمية
90	مذيع أخبار	ياوندى	جمعية الصليب والهلال الأحمر
91	رئيس	أوتاوا	اللجنة الدولية للصليب الأحمر
92	الرئيس التنفيذي	برايا	الاتحاد الدولي لجمعيات الصليب الأحمروالهلال الأحمر
93	مدير تنفيذي	موريتانيا	منظمة خيرية
94	رئيس-مديرعام	انجمينا	جمعية الشارقة الخيرية
95	باحث	سانتياغو	قرى الأطفال إس أوإس
96	باحث ما بعد الدكتوراه	بكين	الجمعية السعودية الخيرية لمرضى الإيدز
97	عالم (صفة)	بوغوتا	منظمةغيرربحية
98	رئيس الوزراء	موروني	مؤسسة محمدالخامس للتضامن
99	حاكم الدولة	كينشاسا	جيش الخلاص
100	رئيس الجمهورية	برازافيل	جمعية المقاصد الخيرية الإسلامية

Table 5. 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