A Combination Method Of Linguistic Features And Machine Learning Techniques For Identifying Arabic Named Entities

Nadia Mohamed Aldali
Department of Computer, Faculty of Education, Elmergib University
Nadiaaldali2@gmail.com

Abstract: Named Entity Recognition (NER) is very significant task in many natural language processing applications, for example; Question Answering (QA), Information Extraction (IE), Text Summarization (TS) and Machine Translation (MT). Many of previous study have addressed the name identification issue in variety of language such as English, Chines and French. But in recently some research efforts have started to focus on Named Entity Recognition for Arabic language. The Machine Learning and Rule-based approaches are popular techniques that are used for Named Entity Recognition to identify and extract named entity such as person name, location name, and organization. In this study, the problem is tacked through integrating three machine learning techniques with linguistic feature in attempt to enhance the overall performance of Arabic Named Entity Recognition. The proposed system is content to two main components including linguistic features, where the linguistic features contents Part Of Speech Tagging (POS), keyword, suffixes and definite article. The machine learning component utilizes the following techniques: Support Vector Machine (SVM), Maximum Entropy (ME) and Conditional Random Fields (CRF) to generate a model for Arabic named entity recognition upon an ANERcorp dataset. The system companied all three machine learning after everyone given the result individually by using voting procedure .Finally the evaluation method that compares the results of the individual machine learning with combination system to compute precision, recall and f-measure. The experiment of this work the best results has achieved when applied the combination system where the results was (92.22%, 95.08%, 94.21%, 91.92% for person, location, organization, and miscellaneous in order). Finally, our system is given the results of high-accuracy in the voting combination system compared with individually methods. And this system is increase in F1 score over baseline for

person, location, organization and miscellaneous named entity compared with previous study.

1: INTRODCTION

Named entity is a term that has been widely used in the field of Natural Language Processing NLP (Benajiba et al., 2008a). The sixth message understanding Conference (MUC-6) has discussed information extraction from unstructured text (1996). One of this information was Named Entity which contains names of persons, organization, locations and numeric expression such as currency, time and date. The objectives of NER that has been flourishing over two decades are classifying and extracting NER features from texts. Nevertheless many of the past studies had concentrated on textual forms and limited fields such as articles and web pages

Different NER applications have different functions. According to (Cowie and Lehnert, 1996) Machine Translating, Information Retrieval, Question and Answering and Text Clustering are examples of the usefulness of the functionalities of NER in Natural Language Processing (NLP) applications.

Machine Translation (MT): The function of MT is translating a document from one language to another. Named Entities (NEs) requires special management so that the text can be translated correctly. Thus, according to (Babych and Hartley, 2003). The quality of the NE translation component can be the part which boosts the overall performance of MT system. When translating from Arabic to Latin languages for instance English, people's names (NEs) are also stated as regular words (non-NEs) in the language without differentiating orthographic characteristics between both the forms. For instance the word "عواطف" "awatf" is used as name of a person and as an adjective meaning "loyalty and truthfulness".

Information Retrieval (IR): According to (Benajiba et al., 2009) information Retrieval (IR) is the chore of finding and recovering appropriate documents from a database of documents based on an input query. IR can benefit from NER in the following ways:-

i) Identifying the NEs in the query.

ii) Identifying the NEs in the documents in order to retrieve the relevant documents by considering their classified NEs.

Question and Answering (QA): this application is closely linked with IR but the results are more complicated. The input for a QA system is the questions and the output is precise and concise answers. NER is exploited to recognize NEs in the questions in order to find the appropriate text and extract the relevant answers (Hamadene et al., 2011), (Friedrich et al., 2006) . As an example, the NE " الشرق الاوسطAlšarq AlÂwsaT "Middle East" can be classified as, a newspaper, a location or ano organisation based on the context. Thus, an appropriate classification for the NE will aid in retrieving the appropriate group of documents which answers the input query.

Text Clustering (TC) :According to (Benajiba et al., 2009) text clustering exploits NER for ranking the resulting clusters according to the ratio of entities which is linked with each cluster This is replicated in improving the process of examining the nature of the clusters and also enhancing the clustering methodology in terms of the chosen features. For instances time expressions along with location NEs can be used as factors which denotes an indication of where and when the events stated in a cluster of documents had occurred.

Characteristics of Arabic Language

In general the main challenge in NLP functions particularly in NER functions is to apply for Arabic documents due to its peculiarity and concise features. Major features of Arabic language which causes a lot of challenges for NER functions are as stated below:-

- **No Capitalization**: in contrast to European languages whereas an NE starts with a capital letter, there is any capitalization in Arabic language. Therefore, capitalization feature is not present in Arabic NER. Thus, the English translations of Arabic words are also listed in the same manner.

- **The Agglutinative Nature**: Arabic is an agglutinative language in which the words may contain lemma, prefixes and suffixes in various combinations which results in a very complex morphology. That mean the word in Arabic language combine one or more of morphological in one word .for example word وبحسناتهم

“wabihasanatihim” this word consist “wa” “and” as a concoction, “bi” “by” as preposition and both as an example for prefix .also in the same word “him” “there” as a possessive pronoun as an example for suffix, And the stem of this word is “hasanat” “virtus”.

- **No Short Vowels:** diacritics or short vowels are required for pronunciation and disambiguation however; most modern Arabic texts do not include diacritics. Hence, an Arabic word might refer to two or more different meaning or words corresponding to the context, it creates too many ambiguity. for example word "رقم" refer to word “number” if is translate as word “raqam” but also can be mean “give” it a number if it translate as a word “rq~am”.

- **Variants in Spelling:** In Arabic, a word with different spelling might still denotes the same word with similar meaning, which creates a many-to-one ambiguity. For instance the word جرام" " “jram” can be written as "غرام" “gram” which has similar meaning.

- **Lack of Linguistic Resources:** The number of Arabic linguistic resources which are available free for research purposes are limited. Those that are available are not suitable for Arabic NER functions because the size of the datasets is not large enough or due to the absence of NEs annotations in the datasets. The Arabic gazetteers are rare and have limited size. Thus, in order to train and assess Arabic NER the researchers have to create their own Arabic linguistic resources.

There are three main important techniques utilized to achieve those two aims are namely the rule-based technique, the ML-based technique and the hybrid technique.

Rule-based technique**:** According to (Mesfar, 2007) rule-based NER systems depend on linguistic rules to identify NEs within the texts using linguistic and contextual clues and indicators. Such systems will exploit dictionaries or gazetteers as auxiliary clues to the rules. The rules are normally carried out in the form of regular expressions or finite state transducers. According to (Meselhi et al., 2014) the maintenance of rule-based systems is not a simple procedure as experienced linguists must be available to provide the system with the proper adjustments. Therefore, any adjustment required for such system is time consuming and labour intensive.

Machine Learning (ML) Approach

As a means to understand the NE tagging method from annotated text, the ML-based NER system exploits the ML algorithms. The machine leaning classify in to categories: supervised learning and semi supervised learning. The different between this two depend to dataset .if that dataset is annotated then we should use supervised learning else we should utilized semi supervised learning, but the most common machine leaning approach is supervised learning. Among others, supervised learning methods (SL) which expresses the NER arrangement task is the standard ML techniques that is often used. It involves a large number of allocated datasets though. Other typical techniques applied for NERs are DT (decision trees), CRF (Conditional Random Fields), HMM (Hidden Markov Models), ME (Maximum Entropy) and Support Vector Machines (SVM).

Hybrid approach

The hybrid method is basically a combination of ML centred method and rule based method that is responsible to optimize the general performances. As such, the flow of the process may begin or end with either ML- based system or the rule based system. Regardless of its practicability in its application, further researches can be done to greatly to enhance its system performances.

2: PROBLEM STATEMENT

In general, Named entities are playing a vital role in many languages thus, extracting such entities is very crucial for several domains such as Sentiment Analysis, Information Retrieval and Machine Translation. Utilizing an appropriate feature for extracting named entity is a key characteristic especially when dealing with complex languages such as Arabic. The effective of extracting named entity in Arabic (ANER) is still low, for that we are trying to enhance the performance of ANER. This could bring many challenges regarding to recognition process. There are several researches that have been presented to extract named entities in Arabic however; there is still demand for improvement in terms of effectiveness. Therefore, this study proposes a combination method between linguistic features and machine learning technique.

3: RESEARCH OBJECTIVES

The objectives of this study are illustrated as follows:

1- To propose a combination method of linguistic approach and machine learning techniques in order to enhance the effectiveness of Arabic NER.
2- To evaluate the proposed method using the common information retrieval metrics Precision, Recall and F-measure.

4: LITERATURE REVIEW

There are several techniques that have been presented regarding to identifying named entities in Arabic for example (Benajiba et al., 2008a), (Benajiba et al., 2008b) presented a variety of methods to recognize Arabic characters in a NER system integrated with machine learning technique. The set consist of lexical, contextual and gazetteer's feature and shall be able to be supported in the Support Vector Machine. Variety of entities in several contexts shall be related to the contextual function. Meanwhile, the lexical feature handles orthographic elements such as special characters, abbreviations, punctuation and digits. The Gazetteer is generally a dictionary list consisting persons, organizations and location names. Classification and type of entities are regulated by SVM. The experiment was conducted with the standard ACE data and a manual UPV corpus.

It is possible to learn the attributes of using both Support Vector Machines and Conditional Random Fields simultaneously. In fact, such studies have already been conducted by (Benajiba et al., 2008b) to using Arabic data set, the morphological, lexical, and contextual features on eight standardized data-sets. The impact and differences obtained were organized respectively and later combined to measure the optimal machine learning technique. ACE 2005, ACE 2004, ACE 2003 sets were used to compile all the results.

Other recent ideas, such as (Meselhi et al., 2014) also adopt the hybrid approach by combing rule based approach that relies on grammatical technicalities. Features are obtained from annotated text by SVM component via machine learning method. This features that are gathered, resembles the morphology, and other elements such as location, person, Organization Experimental. Outcome is based on the Morphological Analysis and Disambiguation and ANERcorp dataset.

5: METHODOLOGY

The framework is split into six primary stages. The first phase discusses the dataset collection and the full details that regarded to it. When the second phase concentrates on the pre-processing, which has two steps such as, remove stop words and tokenization. The third phase features selection to support the dataset for the next step. Then phase fourth is classification by using three approaches which are Conditional Random Filed, Support Vector Machine and Maximum Entropy. The fifth phase is voting. Finally, we will discuss the evaluation based on the Precision, Recall and F-measure in this study of phase sixth.

The figure blow shows the all stages of our system:

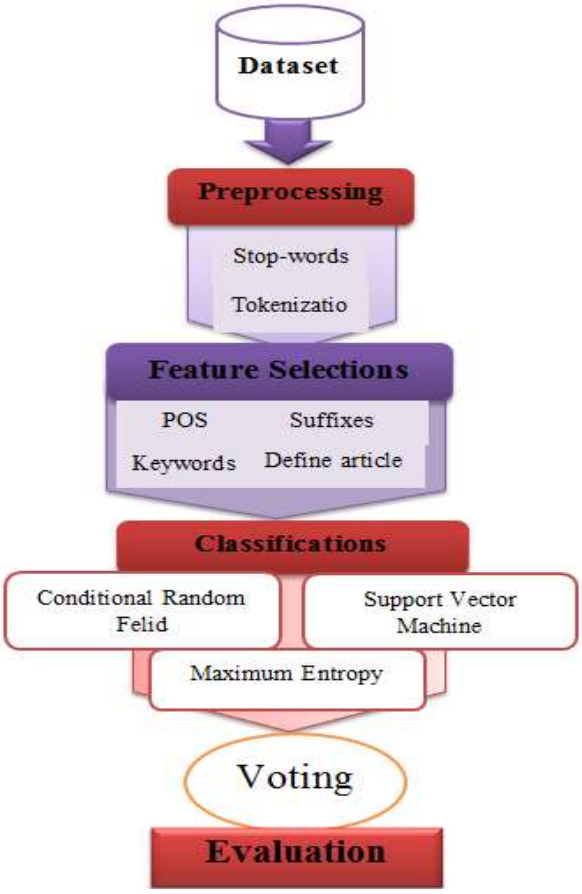

Figure 1. Research Framework

PHASE 1: DATA COLLECTION

ANERCorp is over 150,000 words in the set used in this research. The Arabic corpus annotated manually, which is the incentive to use in the tasks of Arab NER. It consists of two functions; training and testing. There are 11 % of the corpus are

Named Entities, Each token for the corpus is annotated with one of the followings; person, location, organization, miscellaneous (Benajiba et al., 2008a). Table 1 shows gold standard of our data set.

Table 1. Date set gold standard

Class	Total
PER	1887
LOC	1871
ORG	947
MIS	517

PHASE 2: PRE-PROCESSING

Two steps have been executed in this stage such as remove stop words and tokenization to bind the information in a format that can be enabled to recognize the named entities as presented in the next subsections

Stop-word Removed

In this research have been used stop words only from Stanford NLP Arabic stemmer. Stop words are the most frequent words surely are the common words. Anyhow, will list some of the Arabic stop-words, for example the sentence below from the same ANERCorp dataset that used for this study

"وقال خلال مؤتمر صحفي في القاهرة بعد لقائه الرئيس حسني مبارك إن مجلس الأمن قد يتبنى قرارا حول النزاع الأسبوع المقبل. "

Translation to English:

“He said during a news conference in Cairo after meeting with President Hosni Mubarak that the Security Council may adopt a resolution on the dispute next week”

Stop word output: في # بعد # إن # قد # حول "

Tokenization

Tokenization is considered a straight forward procedure that looks for blank spaces and punctuation marks and splits the text accordingly. Example;

Table 2. Tokenization Output

Text in Arabic	*The tokenization output*	*The meaning of text*
على ظهر وجه الارض البسيطة	ala zahr wajh al-ard al-basitah	on the face of the earth
على/ ظهر/ وجة /الارض /البسيطة/	Ala/ zahr/wajh/ al-ard/al-basitah	On/ the/ face/surface/ of /the land/earth

PHASE 3: FEATURE SELECTIONS

The main challenges behind each machine learning techniques based on the features that would be used. Features play an essential role in terms of performance effectiveness of machine learning techniques (Benajiba et al., 2009). There are many types of features that can be used or exploited, for example, semantic features, syntactic features of lexical features or statistical functions. However, each domain of NER requires prescribed features in order to admit entities effectively. For example affixes have a significant impact on biomedical field named entity recognition, therefore most of biomedical entities include affixes (Friedrich et al., 2006). In this work, several features have been used, such as multiple keywords, POS tagging, suffixes and definite articles features. According to the previous studies as mentioned in related work section. These features employed to solving Arabic NER problems by applies a set of features are mostly built from grammatical, syntactic, and orthographic features and a list of keywords. On the other hand the obtained accuracy results were high by using these features which means these features are similar to those highly complicated approaches.

Table 3. Description of used features

Feature	*Role*	
Keyword	Person	This feature aims to identify person's name based on associated keywords

	Organization	This feature aims to identify organization's name based on associated keywords
	Location	This feature aims to identify location's name based on associated keywords
	Miscellaneous	This feature aims to identify Miscellaneous name based on associated keywords
POS tagging	This feature aims to identify the tag of each word	
Suffixes	This feature check if the keyword contains suffix or not	
Definite articles	This feature check if the keyword contains definite article or not	

PHASE 4: CLASSIFICATIONS

The three classifiers SVM, CRF and ME have been applied for Arabic documents with using features in order to evaluate the overall performance of these classifiers for the Arabic document.

i. Maximum entropy (ME)

Maximum Entropy is a type of classifier probably for the exponential model. Maximum Entropy does not consider that it is conditionally independent of each other. The MAXENT is based on the precept of maximum entropy from the entire model that corresponds to training the data; choose one that has the largest entropy. Maximum Entropy classifier can be used to solve a variety of problems such as the detection of text language classification, topic classification, sentiment analysis and more.

$$H(p) = -\sum_{x \in A \times B} p(x) \log p(x)$$

ii. Support Vector Machine (SVM)

Support Vector Machines aims to train data with examples of procedures to optimize the search for the optimal regulation foresees label Invisible example with minimum error rate. It performs the classification using the rules of the examples of training that allows data to identify new instance preciously. Students can be summarized as follows:

$$(\vec{x}_1, y_1), \ldots., (\vec{x}_n, y_n)$$

Where x is the pattern and y_i ⃗_i is the class label. However, there is a formula that needs to be found to train data, here it is:

$$f: X \to \{\pm 1\}$$

Where X is a set pattern that gives x ⃗_i. However, the goal of SVM is to find the optimum hyperplane, so here is a study SVM decision function:

$$f(\vec{x}) = sgn\big((\vec{x} \times \vec{w}) + b\big)$$
$$= \begin{cases} +1: & (\vec{x} \times \vec{w}) + b > 0 \\ -1: & Otherwise \end{cases}$$

iii. Conditional Random Fields (CRF)

Conditional Random Fields (CRF) is one of classifiers in statistical methods used in pattern recognition and machine learning applied to predict structured. While the usual label for the classifier predicts a sample without "neighboring" samples CRF may take into account the context; for example, a linear chain CRF popular in natural language processing sequence predicts label for sequences of input samples.

$$P(\mathbf{A} = \mathbf{y} \mid \mathbf{O}) = \frac{1}{Z(\mathbf{O})} \exp\big(\boldsymbol{\theta} \cdot \mathbf{F}(\mathbf{y}, \mathbf{O})\big)$$

PHASE5: VOTING COMBINATION

The standard method to improve Named Entity Recognition, classification performance is by improving a single classifier. This may involve determining the most informative features, and discarding the uninformative ones (feature selection (Isozaki and Kazawa, 2002)) or finding the correct settings for a specific classifier

(parameter tuning (Tjong Kim Sang and De Meulder, 2003)). An alternative research direction is that of combining several classifiers into an ensemble, and combining their output using a voting procedure (Wang and Kim, 2017). The premise is that mixing a various set of classifiers improves the generalization accuracy, provided that the ensemble's members have sufficient individual performance and their faults do not completely overlap. The yield of the individual classifiers in an ensemble can be combined using the following voting procedure:

Normal majority voting: every classifier casts a vote for a class tag, and the tag with the highest score wins. In case of a tie, the most frequent class is chosen. This is unweighting voting system: all classifiers have an equal measure of influence on the final result of the balloting.

PHASE6: EVALUATION

Evaluation was conducted have been used on this study are the common information retrieval measurement metrics which are Precision, Recall and F-measure. With regard to (Yatskevich, 2003).

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

Precision is defined as the number of correct identified matches compared to the total number of correct matches and false matches identified by the system.

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

Recall is defined as a number of correct identified matches compared to the total of number of correct matches and the needed matches but not identified by the system. Because precision or recall alone cannot accurately evaluate the match quality, so it is necessary to consider a trade-off between them by using the combining measure F-measure which formulated as in Equation.

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5: EXPERIMENTAL RESULTS

The three classifiers of machine learning which are SVM, ME and CRF have been used independently and comparative with the voting combination in order to identify the most accurate class by measuring and comparing their performance.

The Results of the support vector machine.

SVM approach was applied to Training/Testing Arabic data for the Machine Learning system; Table 5 shows the evaluation accuracy for each class of NER.

Tabel 5. Accuracy of each class by using SVM

As

Class	*Total*	*Retrieved*	*Correct*	*False*	*Precision*	*Recall*	*F-measure*
PER	1887	1750	1500	250	85.71%	79.49%	82.49%
LOC	1871	1700	1600	100	94.12%	85.52%	89.61%
ORG	947	750	700	50	93.33%	73.92%	82.50%
MIS	517	400	363	37	90.75%	70.21%	79.17%

clearly shown in Table 5 that the performance of the SVM was very high as it recorded **89.61%** F-measure in class location compared with the other class of named entity.

The Results of the Maximum Entropy.

Maximum Entropy NER approach applied to training and testing Arabic data. Table 6 shows the accuracy of the precision, recall, and F-measure for each class of named entity using the tested system.

Table 6. Accuracy of each class by using ME

Class	*Total*	*Retrieved*	*Correct*	*False*	*Precision*	*Recall*	*F-measure*
PER	1887	1550	1400	150	90.32%	74.19%	81.47%
LOC	1871	1750	1500	250	85.71%	80.17%	82.85%
ORG	947	700	670	30	95.71%	70.75%	81.36%
MIS	517	400	369	31	92.25%	71.37%	80.48%

Table 6 shown that the performance of ME in each class was different and recorded a very high performance 82.85% F-measure in class location compared with the others named entity class.

The Results of the Conditional Random Filed.

The conditional Random Filed approach used to improve the machine learning in Arabic data by training and testing. Table 7 shows the result of evaluation for each class.

Table 7. Accuracy of each class using the CRF

Class	*Total*	*Retrieved*	*Correct*	*False*	*Precision*	*Recall*	*F-measure*
PER	1887	1500	1380	120	92.00%	73.13%	81.49%
LOC	1871	1786	1550	236	86.79%	82.84%	84.77%
ORG	947	787	656	131	83.35%	69.27%	75.66%
MIS	517	380	312	68	82.11%	60.35%	69.57%

Table 7 shown that the performance of CRF recorded 84.77% F-measure in class location which is too high performance compared with the others class.

Results of the Voting Combination with three classifiers.

A voting combination NER approach in Arabic documents was applied to a testing. Table 8 shows the accuracy of the precision, recall, and F-measure for each class of NE (person, location, organization, and miscellaneous) using the tested approach.

Table 8. Accuracy for each class using the combination of three classifications

Class	*Total*	*Retrieved*	*Correct*	*False*	*Precision*	*Recall*	*F-measure*
PER	1887	1800	1700	100	94.44%	90.09%	92.22%
LOC	1871	1810	1750	60	96.69%	93.53%	95.08%
ORG	947	900	870	30	96.67%	91.87%	94.21%
MIS	517	510	472	38	92.55%	91.30%	91.92%

Table 8 shows the comparison between the three standard evaluation measures for classes of NER by applying the combined voting approach. The comparison shows the differences in percentage of each measure in each class. The results reveal that the precision, recall and F-measure in all classes were higher than the previously tested approaches. This proves that the voting combination and use of three classifiers is better than a single approach.

6: RESULT AND DESICCATION

As we see in our system results the combination of the three classifier approaches has shown an enhancement regarding to multiple causes, moreover a set of features that has been used to play an important function in terms of reducing the named entities in limited instances. This has assisted the combination classifiers in order to recognize entities by training. On the other hand, the voting combination has taken the advantages of the three classifiers. This takes in a significant impact on error handling by iterating the three classifiers in a voting mode which guides to handling the incorrect named entity of the first classifier in the second by the third. Moreover, the evaluation of the voting combination approach with the utilized features achieved the highest f-measure comparing with individual approaches

7: CONCLUSION

Arabic language is one of top ten most spoken languages in the world. Addition, there are a huge number of Arabic data published in the internet nowadays. Named entity recognition considered as one of the crucial information extraction tasks in Natural language Processing (NLP) application rely on as an important pre-process step. This work achieved the objective by the evolution and implementation of new systems model NER, to retrieve four types of NER from Arab documents which are Person, Location, Organization and Miscellaneous. The data were collected from ANER Corpus. Moreover, the three classifiers which were SVM, ME, and CRF that have been used in this study have been combined by using a voting combination with the features which were keywords, POS Tagger, suffix and definite article. The study showed that the results are satisfactory and the method used is appropriate.as we are seen in the result the system achieved high accurse and that will be help to enhance overall performance for ANER.

REFRENCES

BABYCH, B. & HARTLEY, A. Improving machine translation quality with automatic named entity recognition. Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT, 2003. Association for Computational Linguistics, 1-8.

BENAJIBA, Y., DIAB, M. & ROSSO, P. Arabic named entity recognition using optimized feature sets. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008a. Association for Computational Linguistics, 284-293.

BENAJIBA, Y., DIAB, M. & ROSSO, P. Arabic named entity recognition: An svm-based approach. Proceedings of 2008 Arab International Conference on Information Technology (ACIT), 2008b. Association of Arab Universities Amman, Jordan, 16-18.

BENAJIBA, Y., DIAB, M. & ROSSO, P. 2009. Arabic named entity recognition: A feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing*, 17**,** 926-934.

COWIE, J. & LEHNERT, W. 1996. Information extraction. *Communications of the ACM*, 39**,** 80-91.

FRIEDRICH, C. M., REVILLION, T., HOFMANN, M. & FLUCK, J. Biomedical and chemical named entity recognition with conditional random fields: The advantage of dictionary features. Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006), 2006. BioMed Central Ltd, London UK, 85-89.

GRISHMAN, R. & SUNDHEIM, B. Message Understanding Conference-6: A Brief History. COLING, 1996. 466-471.

HAMADENE, A., SHAHEEN, M. & BADAWY, O. ARQA: An intelligent Arabic question answering system. Proceedings of Arabic language technology international conference (ALTIC 2011), 2011.

ISOZAKI, H. & KAZAWA, H. Efficient support vector classifiers for named entity recognition. Proceedings of the 19th international conference on Computational linguistics-Volume 1, 2002. Association for Computational Linguistics, 1-7.

MESELHI, M. A., BAKR, H. M. A., ZIEDAN, I. & SHAALAN, K. A novel hybrid approach to arabic named entity recognition. China Workshop on Machine Translation, 2014. Springer, 93-103.

MESFAR, S. Named entity recognition for arabic using syntactic grammars. International Conference on Application of Natural Language to Information Systems, 2007. Springer, 305-316.

TJONG KIM SANG, E. F. & DE MEULDER, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003. Association for Computational Linguistics, 142-147.

WANG, X. & KIM, H.-C. 2017. New Feature Selection Method for Text Categorization. *Journal of information and communication convergence engineering*, 15**,** 53-61.

YATSKEVICH, M. 2003. Preliminary evaluation of schema matching systems. University of Trento.

المستخلص

يعتبر التعرف على أنماط الأسماء ذو أهمية في العديد من تطبيقات معالجة اللغات الطبيعية ، على سبيل المثال ؛ الرد على الأسئلة (AQ)، استخراج المعلومات (IE) ، تلخيص النص (TS) والترجمة الآلية (MT). تناولت العديد من الدراسات السابقة مسألة تحديد الإسم في مجموعة متنوعة من اللغات مثل الإنجليزية والصينية والفرنسية. ولكن في الآونة الأخيرة بدأت بعض الجهود البحثية تركز على التعرف على الكيانات المسماة باللغة العربية وتعد أنظمة التعلم الآلي و منهجية حزمة قاعدة تقنيات شائعة الاستخدام من أجل التعرف على الكيان المسمى واستخراجه مثل أسم الشخص وأسم الموقع والمؤسسة .في هذه الدراسة تم حل المشكلة من خلال دمج ثلاث تقنيات للتعلم الآلي مع ميزة لغوية في محاولة لتعزيز الأداء العام للعترف الكيان باللغة العربية . بالإضافة إلى ذلك ، فإن النظام المقترح يحتوي على مكونين رئيسيان ، بما في ذلك المميزات اللغوية ، حيث تتضمن محتويات الخصائص اللغوية جزءًا من توصيف الكلام(POS) ، والكلمة الرئيسية ، واللاحقة، والمقالة المحددة . ويتضمن مكون التعلم الآلي التقنيات التالية: دعم آلة المتجهات(SVM)، الحد الأقصى لعشوائي النظام (EM) والمجالات العشوائية المشروطة (CRF)لإنشاء نموذج للتعرف على الكيانات باللغة العربية على مجموعة بيانات (ANERcop)، فقد قام النظام بجمع التعلم الآلي بعد أن أعطى الجميع النتيجة بشكل فردي باستخدام إجراءات التصويت . وأخيرًا تم استخدام طريقة التقييم التي قارنت نتائج التعلم الآلي الفردي مع نظام الجمع لحساب معامل الشمولية (F) و الاختبار والدقة ، فإن تجربة هذه الدراسة قد حققت أفضل النتائج عند تطبيق نظام الجمع حيث كانت النتائج (92.22%) ،(95.08%) ، (94.21%) ، (91.92%) للشخص ، والموقع ، والتنظيم ، والممتلكات المتنوعة بالترتيب .وأخيرًا ، أعطى نظامنا نتائج عالية الدقة في نظام الجمع بين التصويت مقارنة بالطرق الفردية .وهذا النظام هو زيادة في درجة معامل F1 على خط الأساس للشخص والموقع والتنظيم والكيان المسمى المتنوع مقارنة مع الدراسات السابقة.