Journal of Engineering and Applied Sciences 13 (10): 3618-3623, 2018
ISSN: 1816-949X

Rule Based Method of Name Entity Recognition for Matching Allah's Finest Names in Holy Quran

[2]Maha M. Hassan, [1]Dhamyaa A. AL-Nasrawi, [1]Redha J. Hassan and [1]Noor T. Mahdi
[1]Department of Computer Science, College of Science, University of Kerbala, Iraq
[2]Department of Computer Science, College of Science (WSCI), University of Babylon, Iraq

Abstract: Named Entity Recognition (NER) is a standout research amongst the most imperative ones in information extraction. Furthermore, it is a sub step in many text processing applications. This study uses rule based method for matching Allah's finest names in specific domain (Holy Qur'an), the proposed system constructs a collection of rules by regular expressions. The proposed rules match all Allah's finest names in Holy Qur'an. Many features were considered in this study including: orthographic, N-grams and affixes. The proposed system solves the problem of appearance Allah's finest names in Holy Qur'an as a substring in another words as well as considers the diacritics (Al-Tashkeel) in Arabic words. Similarity coefficient of names was calculated to reduce the complexity of rules base. The results showed that a rule based by regular expression is an effective, powerful and efficient technique in matching the target names with little collection of rules. This project is a valuable statistical tool that can be right applied to get statistical information about Allah's finest names

Key words: Allah's finest names, Holy Qur'an, named entity recognition, natural language processing, regular expression, rule base

INTRODUCTION

Name entity recognition: The main task of Named Entity Recognition (NER) is to identify and classify essential nouns in a text using a set of predefined categories of interest such as persons, locations, dates, organizations, time, etc. NER play a central role in several applications of Natural Language Processing (NLP) with: machine translation, summarization, question answering or information retrieval (Ismail and Nabil, 2016).

The NER task in a particular language is achieved by collecting knowledge about the language. For instance, in the English language such knowledge may include known titles, capitalization of names, common prefixes or suffixes, Part of Speech (POS) tagging, noun phrases recognition in documents and. In general, procedures that are built for a particular language may not be proper for another language. Many study explored NER problem in a variety of languages and domains. In any case, research that centered around NER for Arabic text is little and limited (Asharef *et al.*, 2012).

The Arabic language and challenges in NER: Arabic is a generally utilized worldwide language that has major features that differentiate it from the other well-known languages, e.g., English and Chinese. Arabic language has assortments of word synonyms, various grammatical forms and different meanings of word depending on features like word order in the sentence. Arabic differs from other languages there is no capitalization, written from right to left and contains 28 characters also diacritics. Arabic language has many forms such as:

- Classical arabic: used in the holy books or pre-Islamic poetry
- Modern Standard Arabic (MSA) which is commonly used in writing, education, media, news, broadcasting, literature and, etc.
- Dialectal arabic (spoken) is the informal day-to-day communication form of Arabic can vary from region to region

As well as, Arabic has vowel marks (Tashkeel or "Harakat" (known as diacritics) (Kanan *et al.*, 2016; Jarrar *et al.*, 2017) (Fig. 1).

Vowel Marks for letters for instance "Alef" are used to identify or distinguish sounds of the letter that are not fully specified by just the letter in Arabic. These characters can be used interchangeably and change the meaning of the word. Then, they are mostly used in the context of verbal exchanges or recitation Tarek (Kanan *et al.*, 2016).

Corresponding Author: Dhamyaa A. AL-Nasrawi, Department of Computer Science, College of Science, University of Kerbala, Iraq

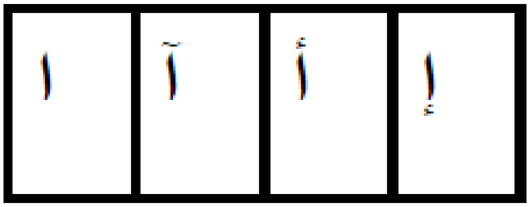


Fig. 1: Diacritics for the letter "Alef"

Arabic block in Unicode (2017) comprising letters, Arabic-Indic digits and diacritics of the Arabic script. The Arabic letters in Unicode range from 0621-063A and from 0641-064A while images and diacritics, range from 0600-06FF.

The NER task is considerably challenging when it targets a morphologically rich language such as Arabic, the most important reasons behind these challenges are.

No capitalization: Arabic does not capitalize nouns.

Agglutination: The agglutinative feature makes it possible for different entities to be concatenated to Named Entity (NE).

Optional short vowels: Short vowels (diacritics) are optional in Arabic.

Inherent ambiguity in named entities: Proper nouns can also represent regular words. For example, the word " حكيم " which means "wise" can be an adjective or a person's name. Another example "ميسلون" (Maysalun) it is both a location name and person's name which makes a conflict for the Arabic NER task.

Spelling variants: An NE can have several transliterations. This leads to many spelling variants of the same word with the same meaning. For example, Person name "Agatha" may yield these spelling variations:" اجاثا" , "اغاتا"or " اغاثا ".

In this study, rule based method of Name Entities Recognition for matching Allah's finest names in Holy Qur'an was demonstrated, the rule based constructed by regular expressions. Several features were considered to make the rules effective, easy and superb these features are: orthographic, N-grams and affixes. The contributions of this study are abbreviated as below:

- Rule based overcomes the appearance of an Allah's finest names in Holy Qur'an as a substring in another words
- Rule based considers the diacritics (Al-Tashkeel) in Arabic words
- Rule based deliberates the similarity of substring in each Allah's finest names with others

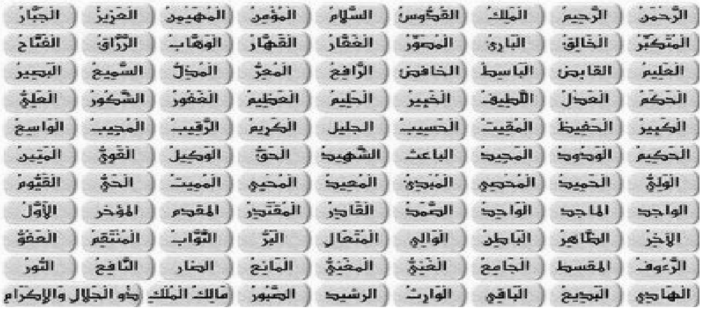
Literature review: Shaalan and Raza (2009) implement a rule based method a names dictionary a local grammar in the formula of regular expressions and a filtering mechanism in NER system for Arabic. The incorrect named entities were rejected by using blacklist which revises the system output. NERA implemented to recognize person, locations, organizations and dates with the following F-measure 87.7, 85.9, 83.15 and 91.6%, respectively.

Abdel-Rahman *et al.* (2010) produce a combination between machine learning techniques to hold ANER problem. The Named Entity (NE) classes are namely person, cell phone, location, car, job, organization, device, currency, date and time classes. Two machine learning techniques were integrated, semi-supervised pattern recognition (namely bootstrapping) and a supervised technique Conditional Random Fields (CRF) classifier. The new approach used pattern and word semantic fields as CRF features and proves their efficiency.

Asharef *et al.* (2012) built an Arabic NER of the crime field. The contributions of his paper are the using of a rule-based NER system to extract and classify NEs from Arabic crime documents. Many features were considered such as information about the surrounding words and their tags, prefix and suffix current word, POS and morphological information and also by utilizing predefined crime and general indicator lists and an Arabic named entity annotation corpus from crime. The result illustrations that the system is effective and the performance of the method is suitable and this is evident from the accuracy of proposed system is 90%.

Shaalan and Oudah (2014) proposed hybrid system that produces state-of-the-art results with an overall 90.66% F-measure on ANERCorp dataset. Alanazi *et al.* (2015) improve a novel Named Entity Recognition (NER) method in the medical domain that use modern Arabic texts to extract names of cancer disease, treatment methods, symptoms and diagnosis methods. A novel system used Bayesian Belief Network (BBN). The results showed that BBN performance is 71.05% overall F-measure. The lowest F-measure score was achieved in recognizing symptoms with 41.66% while the highest was in recognizing disease names with 98.10%.

Zirikly and Diab (2015) studied the influence of word representation and embedding features on Arabic NER performance for Twitter and Dialectal Arabic. A set



الرحمن	الرحيم	الملك	القدوس	السلام	المؤمن	المهيمن	العزيز	الجبار
المتكبر	الخالق	البارئ	المصور	الغفار	القهار	الوهاب	الرزاق	الفتاح
العليم	القابض	الباسط	الخافض	الرافع	المعز	المذل	السميع	البصير
الحكم	العدل	اللطيف	الخبير	الحليم	العظيم	الغفور	الشكور	العلي
الكبير	الحفيظ	المقيت	الحسيب	الجليل	الكريم	الرقيب	المجيب	الواسع
الحكيم	الودود	المجيد	الباعث	الشهيد	الحق	الوكيل	القوي	المتين
الولي	الحميد	المحصي	المبدئ	المعيد	المحيي	المميت	الحي	القيوم
الواجد	الماجد	الواحد	الصمد	القادر	المقتدر	المقدم	المؤخر	الأول
الآخر	الظاهر	الباطن	الوالي	المتعال	البر	التواب	المنتقم	العفو
الرؤوف	المقسط	الجامع	الغني	المغني	المانع	الضار	النافع	النور
الهادي	البديع	الباقي	الوارث	الرشيد	الصبور	مالك الملك	ذو الجلال والإكرام	

Fig. 2: List of Allah's finest names

of features was proposed, comparable performance of proposed system to other systems which use large gazetteers was produced.

Ismail and Nabil (2016) studied the effect of diverse stemming approaches on the Arabic Named Entity Recognition and explored the evidences, restrictions and differences between root-extraction methods and light stemming.

PROPERTIES OF ALLAH'S FINEST NAMES IN THE HOLY QUR'AN

These are the attributes that Allah has given to himself or which he has come to say to his Prophet. God's attributes do not resemble the qualities of creatures. Lack and we conclude from the verse in the verse "Not as His likeness and He is the Hearer, the Seer".

In general, the familiar Allah's finest names are listed in Fig. 1 which used for construction rule based and testing the matching module. Most Allah's finest names have special propriety; they have (AL-Altareef ال التعريف ; Shaalan and Raza, 2009). Which indicates the perfection, assurance, absolute and exclusivity. Various points can be observed (Fig. 2):

- There is an appearance of some names as part of other words, that mean(substring), these names are not considered as one of Allah's finest names such as ("الحي") occurred as substring in ("'الصالحين" "الحياة ")
- Some names do not exist in the Holy Qur'an
- Each name that appeared in the Holy Qur'an without (AL-Altareef) is not considered as one of Allah's Finest Names such as ("'شهيد") but not ("'شهيد"), (" رؤوف') not ("الرؤوف ")
- There is a similarity between some of the Allah's finest names calculated by N-gramm algorithm this similarity was used as one of the features in rule based construction. Such as the similarity of name ("'الباطن") with names (" البارئ"), (الباقي") is 60%

No.	Allah's finest names	No.	Allah's finest names
1	الباسط	20	المجيب
2	الباعث	21	المحصي
3	الباقي	22	المحي
4	البديع	23	المذل
5	الجامع	24	المعز
6	الجليل	25	المعيد
7	الحسيب	26	المغني
8	الحفيظ	27	المقتدر
9	الخافض	28	المقدم
10	الر افع	29	المقسط
11	الر ؤو ف	30	المقيت
12	الشهيد	31	المميت
13	الصبو ر	32	المنتقم
14	الضار	33	المؤخر
15	العفو	34	النافع
16	القابض	35	الهادي
17	الماجد	36	الو اجد
18	المانع	37	الو اسع
19	المبدئ	38	الو الي

Fig. 3: Allah's Finest Names not exist in the Holy Quran

The list of Allah's finest names that are not found in the Holy Qur'an.

METHODOLOGY OF MATCHING ALLAH'S FINEST NAMES

In this study, the methodology of matching Allah's finest names is explained in detail; many steps must be followed for matching process. It mainly consists of four components:

- External list of Allah's finest names
- Selected features
- Construct rule based by regular expressions
- Find all matching

The project passes through four main steps as shown in Fig. 4.

External list of Allah's finest names (Gazetteer): A gazetteer is a list of all the Allah's finest names that used for construction Rule Based and testing the matching process, Fig. 3 shows all Allah's finest names.

Selected features: According to the Arabic text analysis and the Alla's finest names list, the following set of features is selected.

Orthography: Is based on the appearance of the word, e.g., the first letter is a capital letter, all letters are capital or the words consist of digits. In Arabic language, there is no capital letters feature but the short vowels (diacritics) are optional in Arabic words, especially in Classical Arabic form of language that used in the Holy Book.

N-grams: "Grams" means "letters". So, an N-grams is a combination of N letters: a 2-grams is a combination of two letters. N-grams can be used for effective approximate matching. By transforming a arrangement of items to a set of N-grams. This feature is important in constructing the Rule Based because the similarity between names makes it easier to write regular expressions. Each name (string) that is intricate in the comparison process is fragmented into sets of adjacent N-grams. The similarity between two names is accomplished by determining the number of unique N-grams they share and then calculating a similarity coefficient; similarity coefficient is the number of the N-grams in shared (intersection), divided by the total number of N-grams in the two names (union). N-grams works as the following (Algorithm 1):

N-grams algorithm:

```
Inputs: names (words)
Output: Similarity coefficient
1. Split the two words into N-grams
2. Remove the replicated N-grams
3. Arrange the N-grams alphabetically
4. Calculate the shared N-grams in the two terms
5. Calculate the similarity measure
```

In this study, this algorithm is applied for all Allah's finest names. Figure 3 illustrates the some results of similarity between one of Allah's finest names (الباسط) with the rest as example.

Affixes (common word): Some types of NEs share the same word ending or prefix. Morphological features are mainly related to words affixes and roots. Affix feature is a special case of character n-grams where only n-grams at the beginning (resp. end) of the word are used. For instance, words often ends in "er" (reader, writer) or begins in "auto" (automatic, autoplay). In this study, many Allah's Finest Names are shared in substring such as words ends in " يم" " (الكريم , الرحيم , الحكيم) or begins in "الب" (البصير , الباطن) etc.

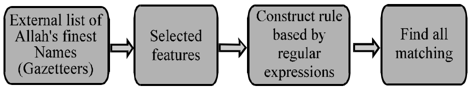

External list of Allah's finest Names (Gazetteers)
Selected features
Construct rule based by regular expressions
Find all matching

Fig. 4: Steps of matching Allah's finest names

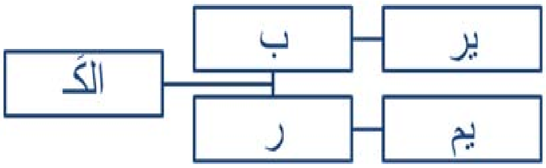


Fig. 5: Names are share the same wordbeginning " الكـ "

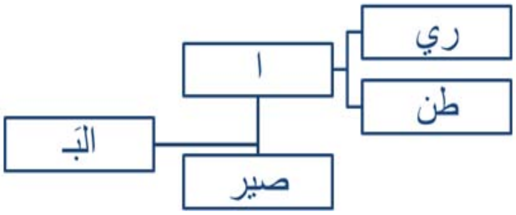


Fig. 6: Names are share the same word beginning " البـ "

Construct rule based by regular expressions: This step focuses on regular expressions which provide powerful, flexible and efficient method for processing text and facilitate the matching of Allah's finest names in the Holy Qur'an. This is a process of the construction rule base of named entity recognition based on regular expressions. The regular expressions depend on the selected features to get a more efficient expression that matching these names in the correct way. The aim of regular expressions is to define strings that match other strings. A regular expression is the same as finding some substring within a document which can be used to match complex patterns of symbols. In this study, the problem of appearance Allah's finest names in Holy Qur'an as a substring in other words was solved using suitable regular expressions. Now, the name (" الحي ") as example was mismatch in the following text: ("المال والبنون زينة الحياة الدنيا"). In addition, the diacritics were considered in writing regular expression. Here, we review some of the regular expressions that were written when the names are sharing the same word beginning (Fig. 5-9).

Thus with the rest of the regular expressions, for example the regular expression matching two names (الكريم , الكبير):

```
الـك [\u0631\u064A\u0645/ \u0628\u064A\u0631]+\s
```

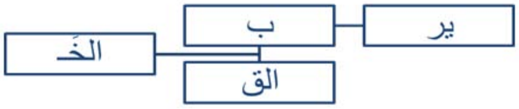


Fig. 7: Names are share the same word beginning “ الخـ ”

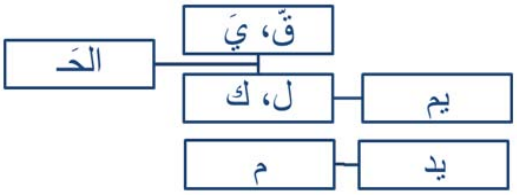


Fig. 8: Names are share the same word beginning “ الحَ ”

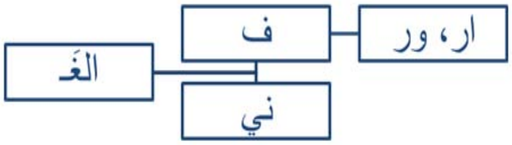


Fig. 9: Names are share the same word beginning “ الغَ ”

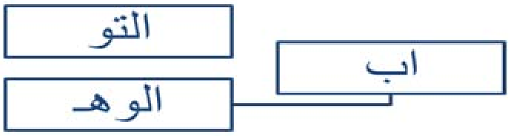
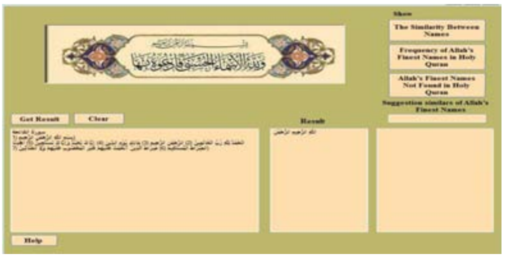


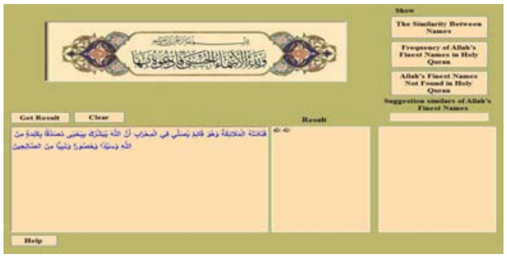
Fig. 10: Names are share the same word ending “ اب ”


بسم الله الرحمن الرحيم
ولله الأسماء الحسنى فادعوه بها
Show
The Similarity Between Names
Frequency of Allah's Finest Names in Holy Quran
Allah's Finest Names Not Found in Holy Quran
Suggestion similars of Allah's Finest Names
Get Result
Clear
Result
Help

سورة الفاتحة
بسم الله الرحمن الرحيم [illegible]
[illegible]

الله الرحيم الرحمن

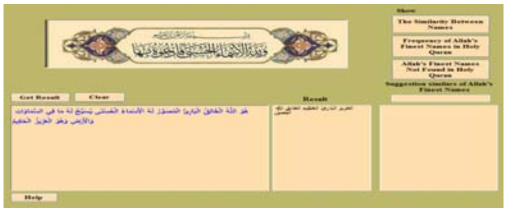
Fig. 11: Result of matching in (Surat Alfateha) سورة الفاتحة


بسم الله الرحمن الرحيم
ولله الأسماء الحسنى فادعوه بها
Show
The Similarity Between Names
Frequency of Allah's Finest Names in Holy Quran
Allah's Finest Names Not Found in Holy Quran
Suggestion similars of Allah's Finest Names
Get Result
Clear
Result
Help

فنادته الملائكة وهو قائم يصلي في المحراب أن الله يبشرك بيحيى مصدقا بكلمة من
الله وسيدا وحصورا ونبيا من الصالحين

الله الله

Fig. 12: Result of matching in other selected holy text


بسم الله الرحمن الرحيم
ولله الأسماء الحسنى فادعوه بها
Show
The Similarity Between Names
Frequency of Allah's Finest Names in Holy Quran
Allah's Finest Names Not Found in Holy Quran
Suggestion similars of Allah's Finest Names
Get Result
Clear
Result
Help

هو الله الخالق البارئ المصور له الأسماء الحسنى يسبح له ما في السماوات
والأرض وهو العزيز الحكيم

[illegible]

Fig. 13: Result of matching on selected holy text

Table 1: Similarity between (الباسط) with the rest

Allah’s finest names	Similarity name	Similar name
الباسط	الباسط	10
الباعث	الباسط	60
الباقي	الباسط	44
البر	الباسط	
البصير	الباسط	40
الواسع	الباسط	36
الحكم	الباسط	22
الواجد	الباسط	20

Like the previous idea, Fig. 10 reviews some of the regular expressions that were written when the names are share the same word ending.

Find all matching in selected text: The result of matching was extracted by collection rules and shown in the holy text of Qur’an. The matching process was implemented to show the result in (Surat Alfateha سورة الفاتحة) which explained in (Fig. 10-13 and Table 1. Now in another selected Holy text.

CONCLUSION

As shown in this study, the accuracy of matching and extracting Allah’s finest names in the Holy Qur’an (specific domain) depends on accuracy of Rule Based in the regular expressions. The selected features (orthographic, N-grams and affixes) are very useful in writing regular expressions as they improve the performance of the program and reduce the number of the rules. This project is a beneficial statistical tool that can be directly applied to get statistical information about Allah’s finest names for example there are 38 names not exist in the Holy Qur’an; the name (الله) has a high frequency in the Holy Qur’an and so on.

RECOMMENDATIONS

For the future researcher, we are planning to consider the meaning and POS (Part of Speech) in a count to improve matching process. In addition, we can use machine learning algorithm with rule based to extract all words in Iraqi dialects in special domain (traditional poetry).

REFERENCES

Abdel-Rahman, S., M. Elarnaoty, M. Magdy and A. Fahmy, 2010. Integrated machine learning techniques for Arabic named entity recognition. Intl. J. Comput. Sci. Issues, 7: 27-36.

Alanazi, S., B. Sharp and C. Stanier, 2015. A named entity recognition system applied to arabic text in the medical domain. Int. J. Comput. Sci. Issues, 12: 109-117.

Asharef, M., N. Omar, M. Albared, Z. Minhui and W. Weiming *et al*., 2012. Arabic named entity recognition in crime documents. J. Theor. Applied Inform. Technol., 44: 1-6.

Ismail, E.B. and L. Nabil, 2016. Exploring the effects of stemming on arabic named entity recognition. Intl. J. Artif. Intell. Appl., 7: 33-43.

Jarrar, M., N. Habash, F. Alrimawi, D. Akra and N. Zalmout, 2017. Curras: An annotated corpus for the Palestinian Arabic dialect. Lang. Resour. Eval., 51: 745-775.

Kanan, T., R. Kanaan, O. Al-Dabbas, G. Kanaan and A. Al-Dahoud *et al*., 2016. Extracting named entities using named entity recognizer for Arabic news articles. Intl. J. Adv. Stud. Comput. Sci. Eng., 5: 78-84.

Shaalan, K. and H. Raza, 2009. NERA: Named entity recognition for Arabic. J. Am. Soc. Inform. Sci. Technol., 60: 1652-1663.

Shaalan, K. and M. Oudah, 2014. A hybrid approach to Arabic named entity recognition. J. Inf. Sci., 40: 67-87.

Unicode, 2017. The unicode consortium. Unicode, Inc., USA. http://www.unicode.org.

Zirikly, A. and M.T. Diab, 2015. Named entity recognition for arabic social media. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, May 31-June 5, 2015, Association for Computational Linguistics, Denver, Colorado, pp: 176-185.