Asian Journal of Information Technology 14 (8-12): 287-293, 2015
ISSN: 1682-3915

A Domain-Based Approach to Extract Arabic Person Names Using N-Grams and Simple Rules

Mohammad Alhawarat
Department of Computer Science, College of Computer Engineering and Sciences,
Prince Sattam Bin Abdulaziz University, P.O. Box 151, 11942 Alkharj, Kingdom of Saudi Arabia

Abstract: Named Entity Recognition (NER) is considered an important task in many human language technologies including information extraction, Natural Language Processing (NLP) and Machine Translation. This is believed to be a challenging task for Arabic language. Most of the existing research studies deal only with names that are found in Modern Standard Arabic (MSA) sources such as news. In this study, we aim at building Classical Arabic name list or Gazetteer which represents an important part of a lively Arabic literature and culture. To achieve this goal, we propose a new approach for extracting Arabic Person Names (APNs). This approach constitutes a new model for extracting named entities from unstructured Arabic text without the need for Part of Speech (POS) tagging and/or morphological analysis. The proposed approach is based on formulating a model that is established on a specific domain. For this study, we use an authentic text in the literature of Islamic-Arabic studies viz, the "Hadith". This domain is related to the Prophet Mohammad's Peace Be Upon Him (PBUH) sayings. To achieve aims of this study, we use NLP and text mining techniques to extract and build an accurate standard list of classical APNs. Also, We built a standard evaluation classical names list in order to evaluate our approach. Results show very good precision of around 84%.

Key words: Entity named recognition, Arabic person names, Natural language processing (NLP), hadith, text mining

INTRODUCTION

Arabic language is considered one of the challenging natural languages due to its properties (Farghaly and Shaalan, 2009). Examples of such properties are: high inflection, derivative and diacritic. Extracting named entities from Arabic text documents is of high importance for many human language technologies such as: machine translation, information retrieval and knowledge extraction. One important named entity is person name, extracting this type of named entities is not as simple as in other languages such English. In English language person names start with capital letters. In Arabic person names has no identifying property as English language and hence is hard to be recognized in a simple manner.

Extracting named entities from specific-domain text documents is important as it serves these domains. For example (Alshref and Aziz, 2014) developed an NER system to extract named entities from political domain. Similarly, authors in (Alanazi *et al*., 2015) developed an NER system to extract named entities from Medical Domain. Also, authors in (Aboaoga and Aziz, 2013) have developed an NER system for sport, economic and politic domains.

The sayings of prophet Mohammad (PBUH) constitute a large resource for classical Arabic text. There are an immense number of his sayings, each saying is said to be a "Hadith". Each Hadith is composed of two parts: "sand" and "matn". The "sanad" part includes a chain of person names who have narrated the Hadith up to prophet Mohammad (PBUH), while the "matn" represents the saying of prophet Mohammad (PBUH). For simple explanation for "Hadith Science" please refer to (Najeeb *et al*., 2015). The dataset that is chosen for this study is very well-known Hadith books in literature; the six books Hadith known in Arabic as "al-Kutub al-Sitta".

Arabic language is considered a low-resource language. Datasets for Arabic language are less than those of Latin languages such as English. Therefore, this study provides a partially standard list for classical APNs for the domain of Hadith. This will be useful for people who work in Arabic human language technologies and hence help in enriching the sources for Arabic language for the field of NLP. To obtain this, we propose an approach for extracting APNs from "al-Kutub, al-Sitta" that is based on domain idea and simple rules. The list that is constructed in this study is considered the first phase of constructing classical APNs list from the sayings of Prophet Mohammad (PBUH).

LITERATURE REVIEW

The researcher in Benajiba have developed a NER system called ANERsys that is based on both N-Grams and maximum entropy. The researchers built their own corpora known as ANERcorp and built their own gazetteers which they called ANERgazet which is built from wikipedia and other internet sources. The system shows good results even though, they did not use morphological analysis. They gazetteer part of person names that they used in their study consists of 2,309 names. They have discussed that using larger gazetteer might improve the system performance as shows.

Elsebai *et al.* (2009), the researchers have developed a person name recognition system for Arabic language. Their system uses a ruled based method that utilizes the Buckwalter Arabic Morphological Analyzer (BAMA), the system is guided by selected keywords to choose the phrases that most likely contain person names. These keywords are one of two types: either introductory verbs or introductory word. Their results outperformed Person Name Recognition for Arabic (PERA) (Shaalan, 2014). The research have extended their work later (Elsebai *et al.*, 2009) where they evaluated their work on 500 news articles that are extracted from Aljazeera website. They've compared their results with (Shaalan, 2014; Mesfar, 2007) although, they are using different data sets.

The researchers in Shalabi have introduced an algorithm to extract proper nouns from Arabic newspaper articles. The algorithm is a rule-based combined with stop words and stemming. The algorithm first removes diacritic and punctuation marks before removing prefixes and suffixes based on predefined lists. After that a set of rules are applied to the stemmed word to check whether its a proper noun. The algorithm is then evaluated using twenty articles that are selected from newspapers.

Extracting person names from modern standard Arabic as well as from colloquial sources have been addressed in Zayed and El-Beltagy. The research claim that their system will extract person names from different types of text sources including news using NLP techniques combined with limited dictionaries.

The research in Zayed have extracted 17,000 person names and family names from the internet and use them as a dictionary. Then patterns around these names are constructed from a downloaded dataset for these names. The pattern building process is achieved by association rule mining approach. Other dictionaries used are stop words and honorifics list. All of these are combined with rules to extract person names from ANERcorp dataset as an evaluation step. Results show that recall is low compared with state of the art approach.

For specific domain NER systems, research have developed a NER system for political domain using morphological analysis combined with rules. In Alanazi *et al.* (2015), the research developed a NER system called NAMERAMA that uses both morphological analysis and Bayesian Belief network to recognize named entities in the context of medical domain.

Most recently, the research in Elbazi and Laachfoubi (2015), built a recognition named entity recognition system called RENA. They claim that their system outperforms the state-of-the-art Arabic NER systems when applied to ANERcorp standard dataset with a F-measure of 93.5%. They used in their system a combination of morphological, lexical, contextual and gazetteers features. For more literature about arabic named entity recognition please refer to (Shaalan, 2014).

In Zaraket and Makhlouta (2012), the research have proposed a system to extract person names and relations from Hadith books. They built an Annotated Narrator Graph Extraction technique (ANGE) from Hadith books. Their system uses morphological analysis, graph algorithms and finite state machines and cross-document reconciliation to extract narrator sequence and then narrator names.

DATA DESCRIPTION AND PREPARATION

There exist large number of Hadith books, therefore, we choose the most well-known Hadith books for nearly all Muslims, viz., the Six books ("al-Kutub, al-Sitta") in order to build an APNs list using the proposed approach. These books are written by well-known narrators and they are: Sahih AlBukhari, Sahih Muslim, Sonan Abi Dawood, Sonan Attermithi, Sonan Annesa'ey and Sonan Ibn Maja. Table 1 shows details of these books. These books are downloaded from in the form of word documents with diacritic.

The preprocessing stage includes converting these word files into plain text documents, then their encoding is converted from UTF-8-CP1256. Diacritic, punctuation marks, numbers and non-Arabic letters are then removed from these text files.

Table 1: Details of the six books used in this study

Books name	No. of Hadiths	No. of words	Unique words	Percentage	Size (MB)
Sahih AlBukhari	7,563	1,004,195	61,086	28.83	6.68
Sahih Muslim	3,033	771,227	52,990	22.14	5.18
Sonan Abi Dawood	5,274	417,720	27,819	11.99	3.08
Sonan Attermithi	3,956	436,077	24,609	12.52	3.19
Sonan Annesa'ey	5,745	427,619	22,817	12.28	3.18
Sonan Ibn Maja	4,341	426,453	33,167	12.24	2.87
Summation	29,912	3,483,291	103,730	100.00	23.61

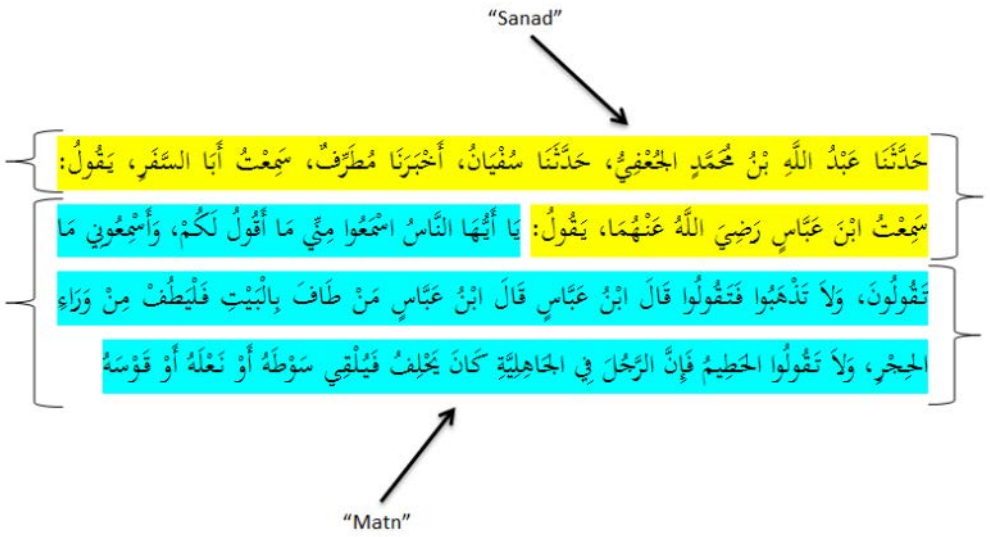

Fig. 1: Normal case where the "matn" follows "sanad"

APNS EXTRACTION

As previously explained, Hadith is divided into two parts: "sanad" and "matn". However, there is no pattern for them, so that they can be separated in a straight forward manner. For example, in some cases the Hadith is divided into exactly two parts following each other: "sanad" then "matn" as Fig. 1 shows. In other cases, "sanad" and "matn" are not two parts, instead they are composed of several parts following each other interchangeably as shown in (Fig. 2).

Note that "sanad" contains, beside narrating words, the names of the narrators which are usually composite names. Composite name in Arabic contains many parts including names and separator words such as "son of" "daughter of","father of", "mother of" and many others. Moreover, the number of narrators is variable.

Therefore, it is not an easy task to separate the "sanad" from "matn". Also, even in some cases where they are separable, still its not an easy task to extract person names as they are mixed with other words inserted by the narrator and sometimes are related to the "matn". Therefore, another approach need to be used to separate "sanad" from "matn". To do so, we propose a methodology to extract "sanad" parts using N-Gram approach. The idea relies on the fact that each sanad part is composed of phrases of narration. These phrases can be formulated as an N-Gram Model where it is composed of phrases starting with a specific words usually called narrating words. Using this method, the text documents are reduced to only phrases begins with one of the narrating words. These narrating words are known by specialized people in the field of Hadith. These words are Domain-specific terms and are extracted manually by analyzing the a group of Hadith Books. These narrating words are shown in Fig. 3. Simultaneously, a list of stop words for Hadith-Domain is constructed that will be used later in the postprocessing for the language model.

One parameter in designing this approach is the number of words in the phrases we are going to extract from the text documents. To capture most of the person names in the "sanad" part and after looking at the documents, 10 g model was a good choice as a starting point.

Fig. 2: One of the cases where the "matn" and "sanad" are not exactly two separable parts (yellow for "sanad" and cyan for "matn")

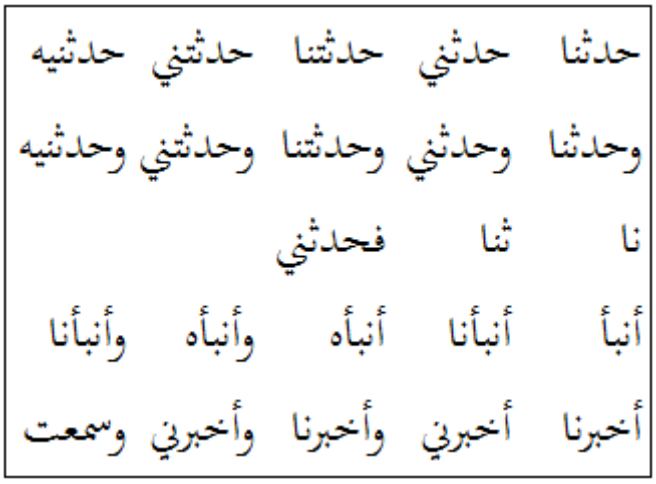

Fig. 3: List of narrating words

APNS EXTRACTION ALGORITHM

In this study, the details of experimentation and evaluation are presented and discussed. This includes extraction algorithm, evaluation of the extracted names and elaborating simple rules to enhance results.

N-Gram phrase extraction: To extract the person names from the six books of the sayings of the Prophet Mohammad (PBUH), the text documents are first preprocessed as explained in section 3, then they are converted into Term Document Matrix in the form of 10 g model as a starting point. After that, The resulted 10 g language model is reduced in size by eliminating any 10 g phrase that doesn't start with one of the narrating words. The original 10 g phrases for all of the six books is 2,174,198, out of them only 76,415 start with one of the narrating words with a percentage of 3.5%. Then, these 10 g that start with the narrating words are considered a new text document and a postprocess of removing the predefined stop words is applied to that text. The resulted list of words is the expected person names of the Hadith dataset. This is composed of 8,064 person names.

After studying the resulted 10 g phrases, it appears that majority of words follow some specific words are not person names, therefore any word follows these words is eliminated. These words are:

Table 2: Summary of the N-Gram phrases

N-Gram	All phrases	Narrating phrases	Percentage
3	1,511,347	11,183	0.74
4	1,748,749	22,648	1.30
5	1,899,824	34,144	1.80
6	2,007,315	45,859	2.28
7	2,085,189	56,213	2.70
8	2,138,813	64,289	3.01
9	2,166,108	70,830	3.27
10	2,174,198	76,008	3.50
11	2,168,412	80,238	3.70
12	2,153,056	83,525	3.88
13	2,130,500	86,007	4.04
14	2,102,436	87,712	4.17
15	2,070,753	88,904	4.29

Table 3: Number of extracted candidate person names

Grams	No. of person names
3	1,421
4	2,154
5	2,851
6	3,407
7	3,930
8	4,509
9	5,124
10	5,796
11	6,453
12	7,264
13	8,123
14	9,044
15	10,045

قال، وقال، قلت، يقول، ويقول، قالت،
وقالت، من، ومن، أن، ولم، أتاه، أتاني

To obtain the best N-Gram that gives the best performance, the same aforementioned procedure that is applied on 10 g is carried out to the range 3-15. This range is used because 2 g model will not catch composite names and 15 is used as it captures long composite names. The number of N-Grams phrases as well as those only start with the narrating words are summarised in Table 2.

The next step is to use the extracted N-Gram narrating phrases and convert them into a Term Document Matrix and then remove narrating words from the resulted Matrix and store the extracted candidates for person names. The number of these extracted candidates from N-Gram phrases is summarized in Table 3.

Evaluation of generated APNs: To evaluate the results above, then we need to compare them with the correct Person Names that is found in the Hadith six-book. To do that manually it will takes a long time to extract person names from documents that are composed of nearly 3,483,291 words. Fortunately, one muslim scholar has already done so by listing the names of all the narrators that is found in the Hadith six-book. This book called Al-Kashif and is composed of numbered items, each represents one of the narrators including his full name, nickname, epithet, his Sir/Madam name if he is a slave, his place of residence, his profession apart from narrating Hadith and those who narrated about him.

First an electronic copy of the book is downloaded from after that the book was studded and analysed in order to find a way to separate the narrator name from other information. The best method was to use the word "about" (in the context of narrating) to split the narrator name from other information. The method achieved high percentage of accuracy and the resulted list is then checked manually to remove irrelevant information. This was a tedious work composed of around three iterations applied on about 7,179 narrator. Then, person names are extracted using Term Document Matrix and filtered by a special list of stop words. The resulted list is 2,877 person, epithet or sure names.

This list is then used to evaluate the N-Gram terms extracted previously. The evaluation is done in terms of recall and precision. Recall, precision and f-measure (also known as F1 score) are well-known measures that are used in information retrieval to evaluate performance of a system. Recall, precision and f-measure are calculated according to Eq. 1-3:

$$\text{Recall} = \frac{\text{Relevententities} \cap \text{Retrievedentities}}{\text{Relevantentities}} \quad (1)$$

$$\text{Pricition} = \frac{\text{Relevent entities} \cap \text{Retrieved entities}}{\text{Relevent entities}} \quad (2)$$

$$\text{Presition} = \frac{\text{Recall} \pm \text{Pricition}}{\text{Recall} \times \text{Pricition}} \quad (3)$$

Where:

Relevant entities = The number of person names correctly extracted from the N-Gram phrases, correctly here means that they exist in the Al-Kashif evaluation list

Retrieved entities = The number of all person names that the model retrieved

It is important here to mention that normalization is applied to some characters of the extracted terms. This is due to the fact that the text documents used in this study are written in different styles according to the approach that is used to convert the original material to electronic form. Therefore, Alef-Hamza might appear in different shapes (أ، إ، آ). Also preceding (و) might be not part of names as it sometimes serves as a conjunction which means "and". Also, the definite article (ال) in most cases is not an original part of names. Therefore, the following normalization procedure is used:

- Any occurrence of (أ،إ،آ) is replaced by (ا)
- Any occurrence of (ا ل ،ال) is eliminated

Table 4: Evaluation of the N-Gram person names according to Al-Kashif evaluation list

Grams	Recall (%)	Precision (%)	F-measure (%)
3	31.60	74.50	44.37
4	44.79	73.48	55.65
5	55.04	72.09	62.42
6	59.55	68.74	63.82
7	62.91	66.77	64.78
8	65.36	64.86	65.11
9	67.18	62.79	64.91
10	69.00	60.94	64.72
11	70.43	59.68	64.61
12	71.76	57.79	64.02
13	73.16	56.50	63.76
14	73.79	55.22	63.17
15	74.25	54.19	62.65

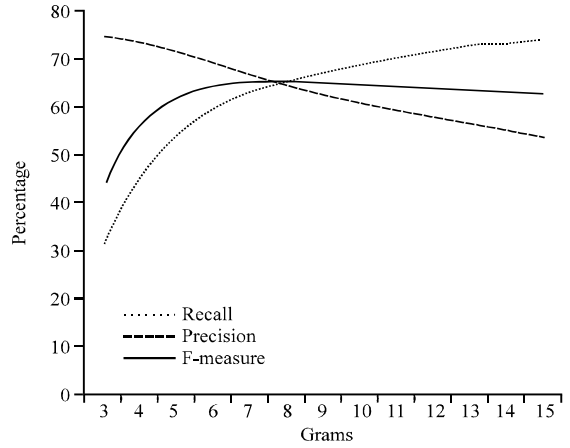

Fig. 4: Performance measures for the N-Gram phrases

- Any occurrence of a preceding (و) is eliminated if the same word without preceding (و) exists either in the evaluation list or in the extracted names list

The recall and precision for all the N-Gram person names are calculated and are shown in Table 4 and also are depicted in Fig. 4.

It is clear from the table that the 15 g attains high recall but its precision is very low. However, the 3 g instance attains high precision with very low recall. In such case the so called F1 score (also known as F measure) would give better indication about the accuracy of the model. Hence, 8 g instance gives best results in this situation. This is might be due to the fact that these gram phrases capture person names in a harmonious manner.

Developing and applying simple rules: Simple rules are developed to avoid non APNs and hence increase precision. These rules are developed based on narrating words and possible sequence of person names. The developed rules are mainly based on the list of words that appear in Fig. 5 and we will call them "Rule Words".

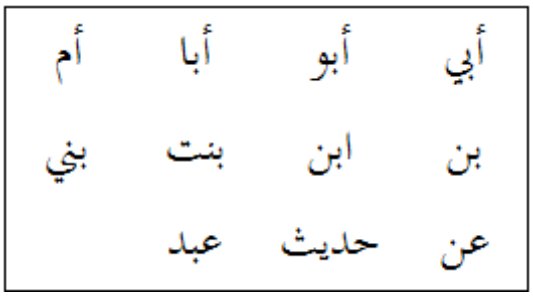

Fig. 5: Words that are used for Rules

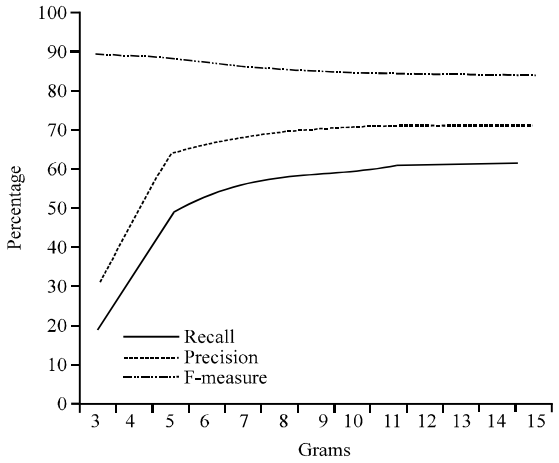

Fig. 6: Performance measures for the N-Gram phrases after applying simple rules

These rules are based on the idea of precedence of one of the "Rule Words", for example if (عن) exists in one of the N-Gram phrases, then the word before it is a person name and the word after it is a person name as well, unless these words are part of "Rule Words" or narrating words (Fig. 3). The same is true for the word (بن). The remaining words have similar rules; if one word appears in the N-Gram phrase, then the word after it is a person name given that it is not part of the words that appear in either Fig. 3 and 5.

One add-on pattern that can be applied to all rules is related to the word next to the adjacent word of the "Rule Words". If that adjacent word starts with the definite article (ال), then it is a person name.

These rules are incorporated in the previous experiments and the performance of the new model is clearly enhanced as Table 5 and Fig. 6.

It is obvious from the results that the 14 g instance gives the best result. Compared to the results before applying the rules, note that the precision is dramatically increased from 64.86-83.84% and the F1-score is increased from 65.11-70.76%. Note also that this enhancement didn't affect the recall much.

Table 5: Evaluation of the N-Gram person names after applying the rules

Grams	Recall (%)	Precision (%)	F-measure (%)
3	18.79	89.05	31.03
4	33.31	87.99	48.32
5	49.30	88.28	63.27
6	53.39	86.85	66.13
7	56.02	86.31	67.94
8	57.52	85.36	68.73
9	58.96	84.67	69.51
10	59.73	84.42	69.96
11	60.43	84.20	70.36
12	60.85	84.09	70.60
13	61.20	83.84	70.75
14	61.34	83.60	70.76
15	61.48	83.31	70.75

CONCLUSION

This study proposes a new approach for extracting APNs from specific domain. The idea is to choose a list of important terms that is specifically, related to the chosen domain and then generate an N-Gram model based on the chosen terms. After that, a tuning phase is needed to choose the best n for that N-Gram Model. The best here means the number that gives the N-Gram Model with the least error when assessed against a prior designed evaluation list of named entities from the chosen domain. Finally, performance might be increased by providing appropriate rules that are mainly related to that specific domain.

In this study, we chose the Hadith domain for two reasons: the first is that it constitutes a wealthy resource of classic arabic names (where other research studies focuses on MSA names), the other is that many scholars need automatic methods to extract information from Hadith books. We develop an algorithm that is used to extract person names based on special terms related to the Hadith domain, viz., the narrating words as shown in Fig. 3. Then, a tuning phase is carried out to choose the best N-Gram Model that maximizes the performance of the model. After that a set of rules are introduced to further reduce the unwanted person names. Results show a vast enhancement in the precision of about 20.

To evaluate our results, we built a standard evaluation list of classical Arabic names based on a simple algorithm that is applied to Al-Kashif book.

RECOMMENDATIONS

The proposed approach could be applied to different domains in the Arabic or other languages for the sake of building standard gazetteers or lists that can be then used in many Human language technologies. Future work may include building a standard Arabic person names list specifically for the domain of Hadith by considering all of the available Books of Hadith (There are hundreds of them).

REFERENCES

Aboaoga, M. and M.J.A. Aziz, 2013. Arabic person names recognition by using a rule based approach. J. Comput. Sci., 9: 922-927.

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.

Alshref, H.H.M. and M.J. Ab Aziz 2014. Named entity recognition for political domain in Arabic language Asian J. Applied Sci., 7: 13-21.

Elsebai, A., F. Meziane and F.Z. BelKredim, 2009. A rule based persons names arabic extraction system. Commun. IBIMA, 11: 53-59.

Farghaly, A. and K. Shaalan, 2009. Arabic natural language processing: challenges and solutions. ACM Trans. Asian Language Inform. Process. Assoc. Comput. Mach., 8: 1-22.

Mesfar, S., 2007. Named entity recognition for Arabic using syntactic grammars. Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, June 27-29, 2007, Paris, France, pp: 305-316.

Najeeb, M., A. Abdelkader, M. Al-Zghoul and A. Osman, 2015. A lexicon for hadith science based on a corpus. Int. J. Comput. Sci. Inf. Technol., 6: 1336-1340.

Shaalan, K., 2014. A survey of arabic named entity recognition and classification. Comput. Linguist., 40: 469-510.