Journal of Computer Science 9 (7): 922-927, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.922.927 Published Online 9 (7) 2013 (http://www.thescipub.com/jcs.toc)

ARABIC PERSON NAMES RECOGNITION BY USING A RULE BASED APPROACH

Mohammed Aboaoga and Mohd Juzaiddin Ab Aziz

Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology, National University of Malaysia, Malaysia

Received 2013-05-13, Revised 2013-06-21; Accepted 2013-06-22

ABSTRACT

Name Entity Recognition is very important task in many natural language processing applications such as; Machine Translation, Question Answering, Information Extraction, Text Summarization, Semantic Applications and Word Sense Disambiguation. Rule-based approach is one of the techniques that are used for named entity recognition to identify the named entities such as a person names, location names and organization names. The recent rule-based methods have been applied to recognize the person names in political domain. They ignored the recognition of other named entity types such as locations and organizations. We have used the rule based approach for recognizing the named entity type (person names) for Arabic. We have developed four rules for identifying the person names depending on the position of name. We have used an in-house Arabic corpus collected from newspaper achieves. The evaluation method that compares the results of the system with the manually annotated text has been applied in order to compute precision, recall and f-measure. In the experiment of this study, the average f-measure for recognizing person names are (92.66, 92.04 and 90.43%) in sport, economic and politic domain respectively. The experimental results showed that our rule-based method achieved the highest f-measure values in sport domain comparing with political and economic domains.

Keywords: Named Entity, Rule-Based Approach, Arabic Morphological Analyzer, Named Entity Recognition

1. INTRODUCTION

The Named Entity Recognition (NER) is very important task in many natural language processing applications such as, machine translation, question answering, information extraction, text summarization, semantic applications and word sense disambiguation. Rule-based approach is one of the techniques that are used for named entity recognition to identify the named entities such as a person names, location names and organization names. It depends on hand-rules that are constructed by the linguistic experts. The main problem of NER is the appearance of named entity in the text which distributes in many types, especially in some languages that are morphologically rich (Alhanini and Aziz, 2011) like Arabic. In rule based approach, the rules that are used for identifying the named entities need the trigger words that reflect the properties of the text in a domain. In spite of some trigger words can be used to identify the named entities in any domains, but others of the trigger word can be used to identify the named entities in one domain only. There are a recent works (Elsebai *et al*., 2009; Shaalan, 2010) that focus on the named entity recognition in Arabic by using the rule based approach, but they developed the rule in political domain and ignored other domains such as economic, sport and health. This means their approach is limited for recognition the political text only and it cannot be used for recognition other domains. Normally, the text in any domain contains all kinds of named entities, for instance person names, company names, city names, country names, sports team and lots of other names from a specific domain. In this study, we introduce four rules for identifying the person

Corresponding Author: Mohammed Aboaoga, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, National University of Malaysia, Malaysia

names in Arabic text. The trigger keywords for person names are presented in three domains: politic, economy and sport. This study is organized as follows. Section 2 introduces the recent works in Arabic named entity recognition using rule-based approach. Section 3 presents the proposed method. Section 4 shows the results and evaluation. Finally, section 5 discusses the conclusions.

1.1. Related Work

This section presents the most related researches on Arabic named entity reorganization that depends on rule based method or machine learning approaches. There are a few works that depend on the trigger keywords and the heuristic rules for recognize the different types of named entity. Maloney and Niv (1998) have presented the technique (TAGARAB) for Arabic name recognition. Their technique consists of two main components: A morphological tokenizer component and a pattern matching engine (name finder). The pre-processing step of Morphological Tokenizer's is to identify the sequences of words, punctuation symbols, numbers that comprise the input text. The Morphological analyzer removes a great deal of morphological ambiguity and has the side-effect of demonstrating that the true difficulties in Arabic morphological ambiguity might be limited to specific contexts. The second component of this technique is the Name Finder module. It uses as input the tokens found by the Morphological Tokenizer with the basic and morphological features attached. It uses data consisting of a set of Pattern-Action rules supported by Word Lists. The latter consists of items such as personal titles that are used by the patterns to recognize names. The Pattern-Action rules use con-textual and structural information about names to recognize them dynamically. They also make extensive use of the feature information coming from the Morphological Tokenizer. They reported that the morphological information is crucially important to effective Arabic name recognition. Abuleil (2004) has presented a rule-based system for extracting the names from Arabic document depending on the handwritten rules and trigger keywords. The main step of his approach is to collect information about the words in the text. After that, the graph will be created in order to represent the relationships between the words. The system was applied on a corpus that contains 500 articles collected from Al-Raya newspaper. He reported that the accuracy of the system equals to 78.4% on the mentioned corpus. Shaalan (2010) developed a system Named Entity Recognition for Arabic (NERA) using a rule based approach. They have presented the results of their attempted at the recognition and extraction of the 10 most important categories of named entities in Arabic script: The person name, location, company, date, time, price, measurement, phone number, ISBN and file name. The resources created are: A White list representing a dictionary of names and a grammar, in the form of regular expressions, which are responsible for recognizing the named entities. A filtration mechanism is used that serves two different purposes: (a) revision of the results from a named entity extractor by using metadata, in terms of a Blacklist or rejecter, about ill-formed named entities and (b) disambiguation of identical or overlap-ping textual matches returned by different name entity extractors to get the correct choice. A person name named entity recognition system for the Arabic Language had been developed and implemented by (Elsebai *et al*., 2009). They have collected the trigger keywords that are used for identifying the phrases that probably include person names. The main component of their approach is the Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) that used for stemming the word in order to overcome the variants of the word that has the same stem. They exploited the General Architecture for Text Engineering (GATE) environment to apply some natural langue processing tasks such as tokenization, sentence splitter, POS tagger, gazetteer, finite state transducer and orthomatcher. The keywords that have been collected in their approach consist of two main lists: Introductory Verb List (IVL) and Introductory Word List (IWL). These lists play an important role in the development of the heuristics and are stored to the GATE system. They introduced the heuristic rule that recognize one type of named entity types: the person name. The rule depends on the position of IVL and IWL words in the text and other words around them. The position is very important to give signal for the rule to start for identification the person name. Finally, they evaluated the rule based method for person name recognition by using the corpus that includes 700 news articles collected from website archives of the Aljazeera.net. The f-measure value that achieved from the system equals to 89%. They reported that their system is better than PERA system. Although this system achieved the reasonable f-measure value, but it cannot deal and recognize other named entity types such as location and organization.

2. MATERIALS AND METHODS

This section describes the rule-based method for Arabic named entity person recognition. The method consists of three main steps: Preprocessing, annotation and the rule application.

2.1. Pre-Processing

This step includes preparing the data, sentence splitter and Tokenization. Preparing the data apply the following procedures (Saif and Aziz, 2011): Remove the extra space between the words; remove all non-Arabic words and symbols from the corpus; normalize the different forms of Arabic letter ("Alef")) (ئ,ء,إ,أ) into the normal form of this letter (ا); delete the diacritics from text .The sentence splitter is the task for segmentation the text into the sentences that compose the text. The main objective of this step is to determine the sentences boundaries in the text in order to segment the text into the sentences. The tokenization is the task that analyses and split the input text into a number of tokens such as, word, number, symbol, space. The main aim of this step is to split the sentence into the tokens that compose it. This step is used the white space between the words to deter-mine the tokens in the sentence.

2.2. Annotations

The main objective of this step is to identify the named entity person names in the text. This step is used the Arabic dictionaries of person names in order to identify this named entity in the text. The identification of the named entity types depends on the dictionary of Arabic named entity types in order to label the 'PERS' on the person name in the text. The dictionary of named entity contains most of the Arabic named entity that are used in natural language. The dictionaries that used in this study had been collected by (Benajiba *et al*., 2007). This step also covers the annotation of keywords and stop-words. The keyword annotation process is to annotate the keywords in the text by using the dictionaries of the keywords that use in the applying the rules. The keywords include the introductory person words list. These keywords have been collected according to the linguistic information about the named entities in Arabic. In this study, the current approach focus on three domains (sport, economic and politic) to recognize the named entities and compare the evaluation results between them. In this step, the keywords have been labeled in the text for each type of these keywords by looking for the keywords that are found in the dictionaries of these keywords. **Table 1** shows the number of trigger keywords in each domain for the named entity type (person name).

The different domains have the different trigger keywords for named entity type (person names) in each domain. For this reason, the trigger keywords in three domains (sport, economic and politic) have been collected in this work. In this step, the trigger keywords for three domains have been collected according to the characteristics of each domain. **Table 2** present some keywords that identify the person name in Arabic text.

2.3. Application the Rules

The main objective of this step is to use the rules in order to recognize the named entities that are not found in the used gazetteers.

In this study, there are four rules for identification of person names in the text that are investigated according to linguistic information of person names. The first rule (**Fig. 1**) is presented here to identify the person names that occur after the Introductory Words Person List (IWPL). IWPL is defined as the words that come before the names as definite word. In relation to that, BAMA is used for getting the Part of Speech (POS) tagging for the words that appear after the IPWL.

As presented in **Fig. 2**, it is crystal clear that the second rule is presented and used to detect the person names that appear before the Introductory Words Person List (IWPL). The person names can be included in the sentence. Hence, their including in the sentence can be used after or before the keywords that are used as a trigger to identify the named entity. So, the person names can be put far away from the keywords. The first and second rule is presented to be used to identify the person names that appear far away from the Introductory Words Person List (IWPL). For example, the following sentence.

" عدن مدينة زار اليمن في الايطالي السفير شتنادير ابرايانم كانمي "

includes the foreign person name (كانمي ابرايانم شتنادير). In this example, the keyword is (السفير) that belongs to IWPL and the Person name (كانمي ابرايانم شتنادير) comes before this keyword. The person name (ابرايانم شتنادير كانمي) is not found in the dictionary of Arabic person names; therefore, the second rule will be applied on this example to recognize the person name. **Table 3** shows the identification of person name in the example.

Regarding to the **Fig. 3**, it is founded that the third rule presented to recognize the person names that occur before the Introductory Verbs Person List (IVPL). To make it clear, the Introductory Verbs Person List (IVPL) is consisted of the verbs which can be positioned before or after the person names.

Table 1. The number of keywords

Domain	Value
Sport	56
Economic	89
Politic	152

Table 2. Some keywords in each domain

Sport	Keyword	Meaning
	اللاعب	player
	المدرب	Coach
	الحكم	governance
	الحارس	The guard
Economic	المحاسب	Accountant
	الصراف	Exchanger money
	الموزع	Distributor
	المدير المالي	Financial Director
Politic	الرئيس	President
	الملك	King
	السفير	Ambassador
	الزعيم	Leader

Table 3. Example of output of rule-based approach

Word	Annotation	Rule based
شتنادير	O	PERS
ابرايانم	O	PERS
كانمي	O	PERS
السفير	IWPL	IWPL
الايطالي	O	O
في	Stop-word	Stop-word
اليمن	LOC	LOC
زار	O	O
مدينة	IWLL	IWLL
عدن	LOC	LOC
Word	Annotation	Rule based
شتنادير	O	PERS
ابرايانم	O	PERS
كانمي	O	PERS
السفير	IWPL	IWPL
الايطالي	O	O
في	Stop-word	Stop-word
اليمن	LOC	LOC
زار	O	O
مدينة	IWLL	IWLL
عدن	LOC	LOC

Table 4. Example of third rule of person name NER

Word	Annotation	Rule based
قال	IVL	IVL
جورج	O	PERS
بوش	O	PERS
ان	Stop-word	Stop-word
الاوضاع	O	O
مستقرة	O	O

```
PerNameR1(S: Sentence)
P=IsIPWL(S) 'Position of Word in S
If P=0 Then
    Exit ' No keyword in S
Else
    For each W in S (From P+1)
      If W ϵ IPWL then
        S= CopyString(S,P+1)
        PerNameR1(S)
      Else
          BMA(W)
          If POS(w)=SW or POS(W)=V Then
            Exit
          Elseif POS(W)=None or POS(W)=PRN Then
                If Pref(W)="ال" and Suf(W)="ي" Then
              Label (W)="Nisbah"
                    Else
                     Label (W)="PERS"
        P=P+1
    End For
End Rule
```

Fig. 1. IWPL rule algorithm

```
PerNameR2(S: Sentence)
P=IsIPWL(S) 'Position of Word in S
If P=0 Then
    Exit ' No keyword in S
Else
 For each W in S (From P-1 to 1)
   BMA(W)
   If POS(w)=SW or POS(W)=V Then
    Exit
   Elseif POS(W)=None or POS(W)=PRN Then
    Label (W)="PERS"
      P=P-1
 End For
End Rule
```

Fig. 2. Algorithm of second rule

For example, the third rule can be applied in the following sentence "مستقرة الاوضاع ان بوش جورج قال". In this example, the keyword is (قال) that belongs to IPVL and the person name (بوش جورج) comes after this keyword. The person name (بوش جورج) is not found in the dictionary of Arabic person names; therefore, the third rule will be applied on this example to recognize the person name. **Table 4** shows the identification of person name in the example.

In the sentence, the person name (علي الحبسي) appear before the introductory person verb (نال) and this rule can be applied to recognize the person name **Table 5**.

```
PerNameR3(S: Sentence)
P=IsIVPL(S) 'Position of Word in S
If P=0 Then
    Exit ' No keyword in S
Else
    For each W in S (From P+1)
      If W ϵ IVPL then
        S= CopyString(S,P+1)
        PerNameR3(S)
      Else
          BMA(W)
          If POS(w)=SW or POS(W)=V Then
            Exit
          Elseif POS(W)=None or POS(W)=PRN
Then
                                Label (W)="PERS"

      P=P+1
```

Fig. 3. Algorithm of third rule

```
PerNameR4(S: Sentence)
P=IsIPVL(S)
If P=0 Then
  Exit ' No keyword in S
Else
    If IPVL ϵ verb-particle Then P=P+1
    For each W in S (From P+1)
      If W ϵ IPVL then
        S= CopyString(S,P+1)
        PerNameR4(S)
      Else
          BMA(W)
        If POS(w)=SW or POS(W)=V Then
            Exit
    Elseif POS(W)=None or POS(W)=PRN
Then
                        Label (W)="PERS"
        P=P+1
      End For
End Rule
```

Fig. 4. Algorithm of fourth rule

Table 5. Example in Arabic the following sentence

Arabic sentence	Meaning
علي الحبسي نال على جائزة أفضل حارس مرمي في أسي	Ali Alhabsi awards the prize of the best player in Asia

As regarded to **Fig. 4**, it is noticed that the forth rule is presented here to recognize the person names that appear before the introductory person verb list (trigger key-words). The introductory person verb list is presented to be included the verb particle constructions that consist of the verb and particle (preposition and adverb). Hence, the detection of the verb in the list is ignored due to its focusing on the person name that comes before IPVL in the sentence.

3. RESULTS

3.1. Dataset

The current corpus is an in-house corpus that has been collected from online Arabic newspaper archives including koora.net, aleqt.net and Alquds.net. This corpus includes three classifications: Sport, economic and politic. It is an electronic corpus of modern standard Arabic that is used for named entity recognition. **Table 6** shows the numerical details about the Arabic corpus used in the method of NER.

3.2. Evaluation

The main objective of this experiment is to evaluate the rule-based approach for recognition the person name in ten documents for each domain: politic, economic and sport. We have used the evaluation method in order to assess our method of named entity recognition. This evaluation method was used to compute the performance measures (precision, recall and f-measure) in the corpus. The method for Arabic named entity recognition has been applied on three corpora of three domains (sport, economic and politic). **Table 7** shows the performance measures (precision, recall and f-measure) for three do-mains of corpus (sport, economic and politic) in the named entity types (person names). It includes the evaluation measures values for the person name in each do-main of the corpus.

Table 6. Statistics on the corpus

Statistic	Value
Size	256KB
Documents	40
Sentences	686
Words	13034

Table 7. The summary of evaluation results

Domains	Precision (%)	Recall (%)	F-measure (%)
Sport	92.57	92.77	92.66
Economic	91.71	92.41	92.04
Politic	92.46	88.57	90.43
All corpus	92.25	91.25	91.71

From **Table 7**, the sport domain has achieved the highest precision, recall and f-measure value (92.57, 92.77 and 92.66%) for the person name by com-paring with the performance measure values for other domains (economic and politic).

4. DISCUSSION

In the evaluation of named entity recognition by using the rule-based approach, the performance measures value of the named entity is affected by the following factors:

- The type of named entity: (the evaluation measures have different values with different named entity types
- The size of corpus: (the corpus with big size may have huge of named entities that is more than in the corpus with small size). The named entities appear with many different positions in the big corpus more than in small corpus
- The number of entities in gazetteer, the number of trigger keywords and rules play an important role in the values of evaluation measures
- The preprocessing: This includes some linguistic tools such as, stop-words detection and morphological analyzer. These linguistic tools are used to disambiguate of some words that need less degree of semantic

5. CONCLUSION

In this study, we have presented the rule-based approach for recognition Arabic named entity. The main aim of this study is to use the rule based approach for recognition the named entities that include person names in economic, politic and sport domain. Our method consists of three main steps: pre-processing, automatic named entity tagged and applying the rules. The method had been applied on Arabic corpus of three domains (politic, economic and sport) to recognize the named entity (person name) in the text. Then, the evaluation method has been used to compute the performance measure for each domain. The experimental results showed that the f-measure value of sport domain (92.66%) is higher than f-measure of other domains: Politic (90.43) and economic (92.04). The technique can be improved to recognize NEs in other domains such as religion and medical and art.

6. REFERENCES

Abuleil, S., 2004. Extracting names from Arabic text for question-answering systems. Proceedings of the Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, (RI' 04), Avignon, France, pp: 638-647.

Alhanini, Y. and M. Aziz, 2011. The enhancement of Arabic stemming by using light stemming and dictionary-based stemming. J. Software Eng. Appli., 4: 522-526. DOI: 10.4236/jsea.2011.49060

Benajiba, Y., P. Rosso and J.M. BenediRuiz, 2007. ANERsys: An arabic named entity recognition system based on maximum entropy. Comput. Linguistics Intell. Text Process., 4394: 143-153. DOI: 10.1007/978-3-540-70939-8_13

Buckwalter, T., 2004. Issues in Arabic orthography and morphology analysis. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, (SL' 04), ACM Press, Stroudsburg, PA., pp: 31-34.

Elsebai, S., F. Meziane and F.Z. BelKredim, 2009. A rule based persons names Arabic extraction system. Commun. IBIMA, 11: 1943-7765.

Maloney, J. and M. Niv, 1998. TAGARAB: A fast, accurate Arabic name recognizer using high-precision morphological analysis. Proceedings of the Workshop on Computational Approaches to Semitic Languages, (SL' 98), Montreal, Canada, pp: 8-15.

Saif, A. and M.J.A. Aziz, 2011. An automatic collocation extraction from Arabic corpus. J. Comput. Sci., 7: 6-11. DOI: 10.3844/jcssp.2011.6.11

Shaalan, K., 2010. Rule-based approach in Arabic natural language processing. Int. J. Inform. Commun. Technol., 3: 11-19.