A Rule Based Persons Names Arabic Extraction System

Ali Elsebai

School of Computing, Science
and Engineering,
University of Salford
Salford M5 4WT, UK

A.Elsebai@pgr.salford.ac.uk

Farid Meziane

School of Computing, Science
and Engineering,
University of Salford
Salford M5 4WT, UK

f.meziane@salford.ac.uk

Fatma Zohra Belkredim

Departement d'Informatique,
Universitie Hassiba Ben Bouali
Chlef, Algeria

fzbelkredim@yahoo.fr

Abstract

Named Entity Extraction is a very new in Arabic Natural Language processing although it has reached maturity for some other languages such as English and French. In this paper, we describe the development and implementation of a person name named entity recognition system for the Arabic Language. We adopt a rule based approach make used of the output produced by the Buckwalter Arabic Morphological Analyser (BAMA), and we used a set of keywords to guide us to the phrases that probably include person names. We have also compared our system with (PERA) Person Name Entity Recognition for Arabic [9] which is based on a lexicon, in the form of gazetteer name lists, and a grammar, in the form of regular expressions. Our system achieves an F-measure of 89% which is an improvement on the results reported by (PERA).

Keywords*:* Message Understanding Conference (MUC), Arabic Named Entity.

1. Introduction

A Named Entity (NE) is the recognition and classification of defined named entities such as organizations (companies, government organisations, committees, etc), persons, locations (cities, countries, rivers, etc) dates and time expressions and monetary amounts (percent, money, weight etc) [1]. The term Named Entity (NE), was first introduced in 1995 by the Message Understanding Conference (MUC-6) [2], and is now widely used and plays a very important role in many areas of Natural Language Processing (NLP) especially in question answering systems, text summarization, text classifications, information retrieval , extraction systems and machine translation [6]. For example, when encountering a proper name in a machine translation application, the system should not attempt to translate it into the target language and a question answering system should not attempt to expand a query containing a proper name [10]. Moreover, names represent a large percentage of unknown words in a text. Furthermore, names are considered as a crucial source of information in a text when extracting contents, clarifying a subject, or identifying related documents in IR systems [11]. Therefore, the accuracy of tools such as chunkiers and parsers in IE systems rely on the recognition of these names.

With the huge amount of published data in Arabic, 200.000 sites and 300.000 users over the net [8], we recognize that developing a system to extract important data from documents becomes essential. However, the Arabic language has its own characteristic and dealing with Arabic language is complicated task. The problem of identifying proper names in Arabic is particularly difficult since they do not start with capital letters so we cannot mark them in the text by just looking at the first letter of the word. Hence, we adopt a rule based approach based on the output results produced by the Buckwalter Arabic Morphological Analyser (BAMA) [4]. Consequently we didn't use any predefined person name gazetteers in our system like the majority of the systems used in the field, and we used a set of keywords to guide us to the phrases that probably include person names.

The remaining of the paper is organized as follows. In section 2, we present the system architecture and describe some of its components. In section 3 we give an example of a heuristic used to identify person names and in section 4 we evaluate our system and compare its results with PERA which is recently developed Named Entity extraction system. In section 5 we conclude and we highlighted our future work

2. System Architecture

The system is composed of two main components. The General Architecture for Text Engineering (GATE) environment [5] and the BAMA. GATE is a language engineering environment developed at the University of Sheffield and has been used extensively for teaching and research since its first release in 1996. There is a set of reusable processing resources provided with GATE, which forms an information system named ANNIE (A Nearly-New IE system) [16]. ANNIE consists of the main processing resources for Information Extraction such as: tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer and orthomatcher. BAMA is widely used in the Arabic Language Processing literature. It has been used in Language Data Consortium (LDC) Arabic POS tagger, Peen Arabic TreeBank, and the Prague Arabic Dependency TreeBank [17]. BAMA is considered as the most respected lexical resource of its kind [14]. In contrast to other morphological analyzers, BAMA performs the input word and returns the stem rather than the root [15]. The word is taken whether it has short vowels or not and the morphological analyzer and the POS tagger are running using the compatible dictionaries and tables. The outcome of this process will be a list of all possible analysis of the input word such as noun, verb, proper noun, adjective, etc. The architecture of the system is summarised in Figure 1 and the various keywords lists developed are described as follows:

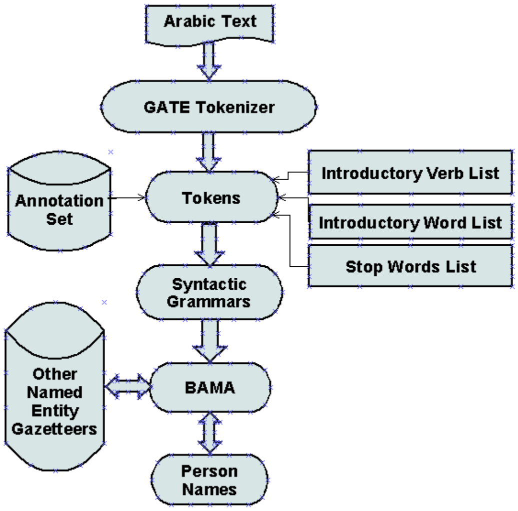

Figure 1: System Architecture

Introductory Verb List (IVL): IVL list contains special verbs that are identified as introducing person names. This list includes verbs such as (said ,قال signed وقع went, ذهب) etc.

Introductory Word List (IWL): IWL list contains a list of descriptives that are identified to be linked to person names. The list contains political functions (Prime Minister, رئيس الوزراء, The President, الرئيس), military titles (General, ,لواء , Commander, قائد), religious title (Imam, إمام, Pope, بابا), Job Titles (Professor, بروفيسور, Doctor, طبيب) etc.

Stop Words: This list is the usual list of words that are not important to the application and include prepositions etc.

Place Names: This list contains the names of known places such as the (Nile River, نهر النيل) etc.

Town and Country Names: This list contains the names of known countries and Towns such as (France, فرنسا, London, لندن Cairo, القاهرة) etc.

Organisation Names: This list contains the names of known organisations such as (the United Nations, الأمم المتحدة The European Union, الإتحاد الأوروبيMicrosoft, ميكروسف) etc.

Arabic Person Names: There are many Arabic names that start with the letters Alif and Lam (AL). These are very often confused with common names as the letters AL are the equivalent of the English definite article (the). As these names are known, the efficiency and precision of the system were largely improved by manually developing a list of all know Arabic names that starts with AL.

IVL and IWL lists play a central role in the development of the heuristics and are added to the GATE system. The text is first used as an input to the GATE system that will perform tokenisation and then annotate the text by highlighting those words that belong to the IVL and IWL lists. We note here that the words in IVL and IWL are not candidate person names but are only used as keywords to find the position of person names in the text. It was relatively easy to define manually the other entities gazetteers as these are limited in number and known. We describe the defined heuristics in section 3.

3. Heuristics Definition

The heuristics are based on the position of IVL and IWL words in the text and other words around them. The application of the heuristics is performed in two phases. The first phase is based on the identification of the position of the verbs from the IVL list. These are performed in the following order:

The algorithm	The Examples
Read word ***w*** from the text IF ***w*** belongs to IVL THEN IF next word belongs to Stop Words THEN find in the text the next word belonging to IVL	أعلن في المؤتمر الأول Announced in the first conference Where the word (في) belongs to Stop Words list

IF ***w*** belongs to IVL
THEN IF next word belongs to IWL
THEN find in the text the next word belonging to IVL
ELSE process word by BAMA
	غادر الرئيس السوري من المطار
	The Syrian president departure from the airport
	Where the word (الرئيس) belongs to IWL

IF ***w*** belongs to IVL
THEN IF next word belongs to IVL
THEN ignore the first word and use the second as a starting point and move to the next word belonging to IVL
ELSE process word by BAMA
	قال السيد محمد خالد
	Mr. Muhammad Kaled said.
	Where the word (السيد) belongs to IVL

Once the text is processed using IVL words, the second step will look at the position of IWL words in the text and the following is performed in this order: IF ***w*** belongs to IWL THEN IF next word belongs to IWL THEN ignore the first word and use the second as a starting point and move to the next word belonging to IWL ELSE process word by BAMA IF ***w*** belongs to IWL THEN IF next word belongs to IVL	نائب الرئيس سليم علي The vice president Saleem Ali Where both words (نائب) and (الرئيس) belong to IWL and the system will ignore the word (نائب) and will consider the word (الرئيس) as the keyword or the start point that must be read.

THEN move to the next word belonging to IWL ELSE process word by BAMA	الملك أكد على أن الإنتخابات ستكون في الشهر القادم The king emphasized that the election will be next month Where the word (الملك) belongs to IWL and the word (أكد) belongs to IVL

IF ***w*** belongs to IWL
THEN IF next word belongs to Stop Words
THEN move to the next word belonging to IWL
ELSE process word by BAMA
	قال السفير إنه يجب علينا مساعدة الدول الفقيرة
	The ambassador said we must help the poor country
	Where the word (السفير) belongs to IWL and the word (إنه) belongs to Stop word

IF ***w*** belongs to IWL
THEN IF next word starts with AL (alif and lam)
THEN PROCESS_AL_WORDS (w)
ELSE process word by BAMA
PROCESS_AL_WORDS (w)
WHILE ***w*** starts with AL
IF ***w*** belongs to list Arabic_Proper_Names
THEN select w as Person name
ELSE ***w*** = next word in text
PROCESS_AL_WORDS (***w***)
	ضابط الشرطة المصري المهدي سالم
	The Egyptian police officer Al-Mahdi Salem
	Where the word (ضابط) belongs to IWL and the words (**الشرطة**) and (**المصري**) will be ignored, then the word (المهدي) will selected as a person name

IF ***w*** is a conjunction and the next word ***w'*** belongs to Country of Place lists then both the conjunction and the next word will ignored THEN process word by BAMA	**محمود عباس إلتقى برئيس الوزراء السابق في لندن توني بلير** Mahmoud Abbas met with the previous Prime Minister Tony Blair in London Where the word (**برئيس**) belongs to IWL and the words (**السابق,الوزراء**) will ignored then the word (**في**) is conjunction and the word (**لندن**) is belongs to Country of Place lists, hence both will ignored
At the end of these two stages all possible proper names are used as an input to the BAMA system that will return all know related words and their classes (verb, noun, proper noun etc.). The following checking is than performed. IF among the words returned by BAMA there is a word ***w*** that is a proper name as shown in Figure 2.	**أعلنت دبي عن الفائزين** Dubai announced the winners Where the word (**أعلنت**) is belongs to IVL although the word (**دبي**) is proper noun as shown in Figure 3, but our system will not mark this word as a person name, because the word (**دبي**) belongs to Town and Country Names lists.

THEN IF ***w*** is not in Countries, Places and Organisations Lists
THEN SELECT w as a person name;
ELSE ignore ***w***	

If the word is known to the BAMA system, then it will return all its classes. However there are cases where BAMA does not recognise a particular word and will not provide a solution as shown in Figure 4 Hence the previous checking rule is extended by the following: IF no solution is provided by the BAMA system THEN select word as person name. In all the cases we have seen so far, this usually points to a non Arabic Proper Name.	**رئيس فريق المفاوضات الإيرانية لاريجاني** The Iranian negotiation team leader Larigany Where the sequence of words (**رئيس فريق**) belongs to IWL and both words (**المفاوضات**, **الإيرانية**) will ignored then the word (**لاريجاني**) will process by BAMA, although the system not provide any solution as shows in Figure 4, the word selected as a person name

```
Initializing in-memory dictionary handler.
Loading dictionary : dictPrefixes .
78 entries totalizing 299 forms
Loading dictionary : dictStems ..................
38600 lemmas and 47261 entries totalizing
82158 forms
Loading dictionary : dictSuffixes ..
206 entries totalizing 618 forms
Loading compatibility table : tableAB ...
1648 entries
Loading compatibility table : tableAC .
598 entries
Loading compatibility table : tableBC ..
1285 entries
... done.
Initializing in-memory solutions handler.
... done.
possible analysis
of the input word سالم =VERB_IMPERFECT
possible analysis
of the input word سالم =NOUN_PROP
```

Figure 2: Buckwalter output for the word (Salem, **سالم**).

```
Initializing in-memory solutions handler.
... done.
possible analysis
of the input word دبي =NOUN
possible analysis
of the input word دبي =ADJ
possible analysis
of the input word دبي =NOUN
possible analysis
of the input word دبي =NOUN_PROP
```

Figure 3: Buckwalter output for the word (Dubai, دبي)

```
Loading compatibility table : tableAC .
598 entries
Loading compatibility table : tableBC ..
1285 entries
... done.
Initializing in-memory solutions handler...
... done.
possible analysis
of the input word لارجاني =  No Solution
```

Figure 4: Buckwalter output for the word (Larigany, لارجاني)

4. System Evaluation

We evaluated our system using around 700 news articles extracted from the Aljazeera television website [3] and we compared our system with PERA by both including and excluding the gazetteers. The results obtained show that our system performs significantly better PERA. The results are summarised in Table 1. However our system differs from PERA on the following aspects:

the arrangement of the First, as seen above Arabic phrase does not always take one state. Sometimes the proper noun in the phrase appears next to the keyword and sometimes appears after four or five words after the keyword and sometimes the proper noun appears before the keyword and sometimes the proper noun is completely omitted from the phrase. Consequently we can’t constantly mark the words next to the keyword as a proper noun, as **الملك الأردني عبد الله الثاني**this example shows [9]
The Jordanian king Abdullah II,

This phrase can exist in different form in the Arabic text as shows in Figure 5. However the rules defined in PERA can only handle this state of the phrase and lacks the ability to deal with the other forms which can be come up while processing the text.

الملك الأردني السابق المتوفى في سنة 1999 حسين بن طلال
The previous Jordanian king Husain Bin Talal
died in 1999

غادر عبد الله الثاني الملك الأردني من عمان إلى لندن
Abdullah II the Jordanian king departed
from Amman to London

قال الملك الأردني إن المؤتمر سيعقد في نوفمبر القادم
The Jordanian king said the conference will
take place next November

Figure 5: different forms of Arabic phrase

As `we illustrated above our system can deal with all these kind of phrases. Secondly, contrary to the PERA system, our system rely on the output results given by BAMA and thus does not require any predefined person name gazetteers. A likely listing a huge amount of entries (472617 used in PERA) in several gazetteers decreases the analysis speed of the system.

	Our system	PERA with gazetteers	PERA without gazetteers
Precision	93	85.5	80
Recall	86	89	70
F-Measure	89	87.5	75

Table1: System Evaluation

5. Conclusions and Future work

In this paper, we reported the early stages of our rule based Arabic person name extraction system which makes use of the GATE and BAMA systems. In contrast with the majority of the systems used in the field, we did not use any predefined person name gazetteers. We have also compared our system with the PERA system a recently developed system. Our system achieves an F-measures of 89% and shows that the results are significantly better than those reported by the PERA system. Some possible future work includes extracting the rest of the named entities such as organization, location, date, etc. Moreover we will compare our system with other existing systems.

References

[1] Gaizauskas R, Wilks Y “Information Extraction: Beyond Document Retrieval”, Memoranda in Computer and Cognitive Science, 1997, CS-97-10.

[2] Chinchor N, “MUC-7 Named Entity Recognition Task Definition”, Version 3.5, http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html, 1997.

[3] Aljazeera TV, http://www.aljazeera.net/, 2008

[4] Buckwalter T, “Buckwalter Arabic Morphological Analyzer”, version 2.0. LDC catalog No LDC2004L02, Linguistic Data consortium, 2004, www.ldc.upenn.edu/Catalog.

[5] Maynard D, Cunningham H., Bontcheva K, Catizone R., Demetriou G, Gaizauskas R, Hamza O, Hepple M, Herring P, Mitchell B., Oakes M., Peters W., Setzer A., Stevenson M., Tablan V., Ursu C. and Wilks Y, “ A Survey of Uses of GATE” , *Technical Report CS-00-06*, Department of Computer Science, University of Sheffield, 2000.

[6] Cowie, J., and Lehnert W, “Information Extraction”, *CACM*, 39(1): 83-92, 1996.

[7] Ibn Auda, D., 2003. Period Arabic Names and Naming Practices, *In Proceedings of the Known*

[8] Mesfar, S, “Named Entity Recognition for Arabic using syntactic grammars”, *Proceedings of the 12th International Conference on Application of Natural Language to Information Systems,* 2007, pp 305-316, Paris, France.

[9] Shaalan, K. and Raza, H, “Person Name Entity Recognition for Arabic”, *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic, 2007 pp. 17-24, June.

[10] Wacholder N., Ravin Y., and Choi M, “Disambiguation of proper names in text”, *In Proceedings of the 5th Applied Natural Language Processing Conference,* , pp 202–208, Washington, D.C., March, 1997

[11] Rau, L. F, “Extracting Company Names from Text”, Proceedings of the Seventh Conference on Artificial Intelligence Applications, Feb. 24-28, Miami Beach, Florida, pp.29-32, 1991

[12] L. Eikvil, “Information extraction from world wide web - a survey”,Technical Report No 945, ISBN 82-539-0429-0, Norweigan Computing Center, Norway, 1999.

[13] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, “Performance measures for information extraction”, In Proceedings of the DARPA Broadcast NewsWorkshop, Herndon, 1999 VA, Feb.

[14] Hajic J, S. O., Buckwalter T, Jin H, . Feature-Based Tagger of Approximations of Functional Arabic Morphology. The Fourth Workshop on Treebanks and Linguistic Theories. Universitat de Barcelona, 2005..

[15] Larkey, Leah S. , Ballesteros, Lisa, and Connell, Margaret E., Light Stemming for Arabic Information Retrieval: Knowledge-based and Empirical Methods, A.Soudi, A. van den Bosch, and Neumann, G., Editors. Kluwer/Springer's series on Text, Speech, and Language Technology. 2007

[16] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL’02), 2002.

[17] Attia M., An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks, The Challenge of Arabic for NLP/MT Conference. The British Computer Society, London, UK., 2006.