Proper Noun Extracting Algorithm for Arabic Language

Riyad Al-Shalabi, Ghassan Kanaan, Bashar Al-Sarayreh, Khalid Khanfar,
Ali Al-Ghonmein, Hamed Talhouni, and Salem Al-Azazmeh

Arab Academy for Banking and Financial Sciences, Jordan
rshalabi@aabfs.org, ghkanaan@aabfs.org, bsarayreh@aabfs.org, kkhanfar@aabfs.org

Abstract

Many of Natural Language Processing (NLP) techniques have been used in Information Retrieval, the results is not encouraging. Proper names are problematic for cross language information retrieval (CLIR), detecting and extracting proper noun in Arabic language is a primary key for improving the effectiveness of the system. The value of information in the text usually is determined by proper nouns of people, places, and organizations, to collect this information it should be detected first. The proper nouns in Arabic language do not start with capital letter as in many other languages such as English language so special treatment is required to find them in a text. Little research has been conducted in this area; most efforts have been based on a number of heuristic rules used to find proper nouns in the text. In this research we use a new technique to retrieve proper nouns from the Arabic text by using set of keywords and particular rules to represent the words that might form a proper noun and the relationships between them.

To extract proper nouns from the retrieved document, we need some information about it and where it was found. First, we mark the phrases that might include proper nouns; second, we apply rules to find the proper noun and we use simple methods (stop wording and stemming) usually yield significant improvements. To test the system we have used 20 articles extracted from the Al-Raya newspaper published in Qatar and Alrai newspaper published in Jordan.

Keywords- Proper noun, Arabic Language, Prefixes, suffixes.

1. INTRODUCTION

The core of information retrieval task is to find and retrieve documents relevant to given query from collections, generally where query and documents are in the same language. Several other IR tasks use very similar techniques, e.g. document clustering, filtering, new event detection, and link detection, and they can be combined with NLP in a way similar to document retrieval. Recent research has extended this goal to include document collections in languages different from the language of the query, known as Cross-Language Information Retrieval (CLIR) [1].In information retrieval, proper nouns in queries frequently serve as the most important key terms for identifying relevant documents in text. [2].

Arabic language is currently the sixth most widely spoken language in the world. It is the mother tongue of about 300 million of peoples [3]. Arabic is an official language in more than 22 countries. Since it is also the language of religious instruction in Islam, many more speakers have at least a passive knowledge of the language. The direction of writing is from right-to left, and the Arabic alphabet consists of 28 letters. As discussed in [4], the Arabic alphabet can be extended to ninety elements by writing additional marks, vowels, and different shapes according to their position in the word. Most Arabic words are morphologically derived from a list of roots; most of these roots are three

constants.

The Arabic language differs from other natural languages such as English language, its own features that are not found in other languages. Natural Language Processing (NLP) in the Arabic language is still in its initial stage compared to the work in the English language, which has already benefited from the extensive research in this area. There are some aspects that slow down progress in Arabic Natural Language Processing (NLP) compared to the accomplishments in English and European languages [5].

These aspects include:

- The absence of diacritics in the written text creates ambiguity and therefore, complex morphological rules are required to identify the tokens and parse the text.
- The direction of the writing of the script is from right to left and some of the characters change their shapes based on their location in the word.
- Capital letters are not used in Arabic, which makes it hard to identify proper names, abbreviations., and this creates increased ambiguity and especially complicates such tasks as Information Extraction in general and Named Entity Recognition in particular.
- The major difference is that Arabic is mainly highly inflectional and derivational, which makes morphological analysis a very complex task while English and other languages are concatenate [6].

In addition to the above linguistic issues, there is also a lack of Arabic corpora, lexicons, and machine-readable dictionaries, which are essential to advance research in different areas.

It is important to note the difference between prefixes and suffixes functions in Arabic and their functions in other languages. In modern English and modern French for example, the prefix or the suffix is usually a modifier of the meaning of the noun or the verb. It does not add any entity (happy prefixes with un _ unhappy). In Arabic, the prefix can add an entity to a noun or a verb. For example, the prefix can be a preposition and the suffix can be a pronoun. Figure 1 tells more about prefixes and suffixes in Arabic [8][23][24][27].

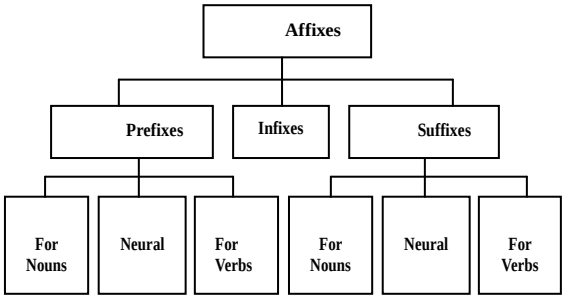

Figure 1: Classification of Affixes

The stemming algorithm is a computational process that gathers all words that share the same stem and have some semantic relation [3]. The main objective of the stemming process is to remove all possible affixes and thus reduce the word to its stem. It is normally used for document matching and classification by using it to convert all likely forms of a word in the input document to the form in a reference document [9].

Arabic stemming algorithms can be classified, according to the desired level of analysis, as either stem-based or root-based algorithms. Stem-based algorithms, remove prefixes and suffixes from Arabic words, while root-based algorithms reduce stems to roots [10]. Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without trying to deal with infixes or recognize patterns and find roots [11]. Al-Shalabi developed a system that detects the root and the pattern of Arabic words with verbal roots [12]. Al-Jlayl and Frieder showed that stem-based retrieval is more effective than root-based retrieval [13].

2. ARABIC WORDS

The two English words noun and name are both translated into Arabic by Ism ( ). A “name” in English is considered to a subclass of a noun referred to as proper nouns, which is also true for Ism in Arabic. Ism is one of the three major part of speech categories in the Arabic language i.e. nouns, verbs and particles (in Arabic, Ism ( ), Fi’l ( ) and Harf ( ) respectively) , figure 2 shows major part of speech categories.

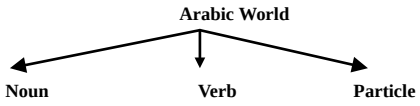

Figure 2: A Classification of Arabic Words According to The Part of Speech

Particle in Arabic is voice-based segment of excerpts of throat or tongue or lips. Such as: on, in, of ((. The Particle class include: prepositions, adverbs, Conjunctions, and interjections.

Verb is a word that indicates an action or state with being connected with notion of time. Verb is divided into three Classes: Past tense , present tense , and ordered tense , such as: ( ).

Noun or ism is a word that indicates meaning by itself without being connected with the notion of time, and that describes a person, location, or idea. Such as (Ali, Maca, and Bird), in Arabic ( ).

There are two main kinds of nouns: variable and invariable. Variable nouns have different forms for the singular, dual, plural, masculine, feminine, diminutive, and relative. Variable nouns: some of them are fixed (solid) nouns and some of them are derived; fixed noun: The fixed noun is not derived from another word, i.e., it does not refer to a verbal root. And derived nouns: these are nouns that are built according to the Arabic derivation rules. We refer to these as proper nouns in this paper, but it should be understood that this usage is not restricted to names of people (personal name) [14] [5][27]. Figure 3 shows examples of proper nouns.

Name (Ism) is sub classified into three sub categories: Alam “Proper noun”, Masdar “infinitive”, and Sifah “adjective" [23][24] [26][27].

Division proper noun, as the word to a single, and composite:

- Singular proper noun: is a proper noun, consisting of one word. Such as: Mohammed, Ahmed, Ali and Ibrahim, Suad, Khadija, Mariam, and India.
- Composite proper nouns: is a proper noun that consisting of two or more, and shows one fact before and after transport. Such as: Abdullah, Abdul Rahman, Abdel Mawla, and some of them are Kunia: Abu Bakr, Abu Obeida, Abu Ishaq, Abu Jaafar.

Our proper noun classification, which was developed through corpus analysis of newspaper texts, is organized as a hierarchy which consists of 7 branching nodes and 20 terminal nodes. Currently, we use the names of people (person) , places (location) , organizations , things , ideas , events , dates , times , or other entities to assign categories to proper nouns in texts [16]. Figure 3 shows a hierarchical view of our proper noun categorization.

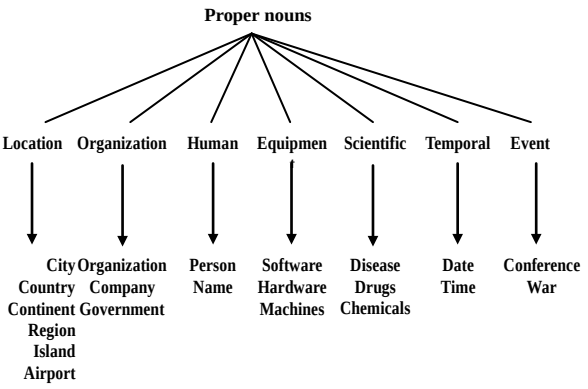

Figure 3: Proper Noun Categorization

3. ARABIC NAMES

Arabic names usually consist of the designation long. They do not consist of first name, middle name and surname, but also of a long series of names, a system used in the whole of the Arab world. Given the importance of the Arabic language in Islam, he uses the vast majority of Muslims around the world Arabic names. But rarely used as a label run outside the Arab world. Figure 4 shows structures of Arabic names.

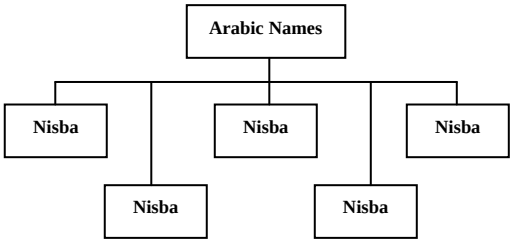

Figure 4: Structure of Arabic Names

- **Ism** name is a means to define a specific person, his or her personal name (e.g. "Ali" or "Fatima"), and often the meaning of the Arabic names be returned refers to a benign such as "Samir" means "friend" or "Kareem " means "generous", and both words are employed as adjectives and nouns in regular language , and tend Arab identity of naming names have a religious reference, such as "Muhammad" or " Yousef "or" Abdul Rahman. The Ism are divided into: Ism consisting of one part such as ("Salem" , "Hamed") ,and Ism consisting of two parts such as) Abdul Wahab , Ezzedine). Arab newspapers sometimes try to avoid confusion by placing names in brackets or between quotation marks. Generally, context and grammar will indicate how the word is being used, but foreign students of Arabic may initially have trouble with this. A very common form for Muslim Arab names is the combination of **Abd** followed by often one of the Muslim 99 Names of God (e.g. Abdullah).To an extent most Christian Arabs do not use specifically Muslim names such as (Mohammad (. There are also Arabic versions of Christian names, and names of Greek, or Armenian, Adoption of European names (e.g. George).
- **Kunya** Kunya (Nickname) is a means to define the person by the first son or daughter is the first by the addition of the word **"Abu" ,Aba or "Aby"** as the store's name at the beginning of a Bedouin boy or girl. Often, a kunya referring to the person's first-born son is used as a substitute for the ism ("Abu "): (e.g "Abu Karim") for "Father of Karim". It can refer to the person's first-born daughter. The female variant is **("Umm")**, thus ("Umm Karim"(. Sometimes required to begin a wordy following: Father, mother, son, daughter, brother, sister, and his uncle, uncle. About: Abu Khaled, Umm Yousef, and alwaleed son, the daughter of Zaid Ansariyeh, Baker's brother, and sister-Ansar, his uncle Ali, and uncle Yusuf. In Arabic in order:

. :

.

- **Nasab** The nasab is a patronymic or series of patronymics. It indicates the person's heritage by the word **(Ibn)** sometimes **(bin)** which means "son". Thus (Ibn Khaldun) means "son of Khaldun" (Khaldun is the father's ism (proper name)).The Arabic for "daughter of" is (Bint) A woman with the name "Fatimah bint Ahmad bin Haroun" translates as "Fatimah, daughter of Ahmad, son of Haroun".
- **Laqab** The laqab is intended as a description of the person. So, for example, in the name of the famous Abbasid Caliph Haroun al-Rashid (of A Thousand and One Nights fame), Haroun is the Arabic form for Aaron, and "al-Rashid" means "the righteous" or "the rightly-guided".
- **Nisba** The nisba describes a person's occupation, geographic home area, or descent (tribe, family, etc). It will follow a

family through several generations, and it is for examples common to find people with the name Al-Ordoni (the Jordanian, or rather "of Jordan"), and Al-Misri (the Egyptian, or rather "of Egypt") in many places in the Middle East, despite the fact that their families may have resided outside Jordan or Egypt for several generations. The nisba, among the components of the Arabic name perhaps most closely resembles the Western surname and sometimes become family of person [22][27].

4. ARABIC NAME DETECTION

Identifying proper noun in Arabic is particularly difficult, since names in the Arabic language do not start with capital letters so we cannot mark them in the text by looking at the first letter of the word. There is no fixed method to name in the Arabic language, there are multiple ways of writing the name ; for example, frequently use the word **"Ould" ( )** that means "son of" in some North African countries such as Mauritania," Mauritanian poet Ahmed Ould Abdul Kader" in Arabic " ". While spreading the use of the word **"bin or ibn"** that means "son of" in some of the Middle East and Arab Gulf countries, as the method to name in the old Arab Islamic name, such as Prince Mohammed bin Rashid Al Maktoum.

Modern naming convention may drop the words "bin","ibn", "ould", or "bint" as it is already implied, which showed ratios son to his father in many Arab countries, so Fatimah's full name would be "Fatimah Ahmad Haroun Al fulany"

.

In this paper first we use previous structure of Arabic names to guide us to mark Arabic name in text, second we use set of keywords that help us to identify and detect place of Arabic Names, where we can find them in the text and extracts them from the text, this keyword usually followed by a personal name. **Abd X** means slave of X where X is a word describing Allah (God) (e.g. Abdul aziz). **Abu** means father of Y, **Umm** means mother of Y, **Ibn** or **bin** means son of Y where Y is personal name . In such a case the personal name would be prefixed to bin or Ibn .

Abu Karim Muhammad al-Jamil ibn Nidal ibn Abdulaziz al-Filistini

"Father-of-Karim, Muhammad, the beautiful, son of Nidal, son of Abdulaziz, the Palestinian" (karim means generous, Muhammad means praised Jamil means beautiful; Aziz means Magnificent, and it is one of the 99 names of God) .Abu Karim is a kunya, Muhammad is the person's proper name (ism), al-Jamil is a laqab, Nidal is his father (a nasab), Abdulaziz his grandfather (second-generation nasab) and "al-Filistini" is his family nisba.

If the person has performed the (Hajj), the honorific **(“Haji")** would be prefixed to his name, (e.g. Haji Muhammad (. Another words that prefix the person name **("Mr." or "Sheikh") ("Sharifah" , "Mrs." for females).**

5. DETECTION AND EXTRACTING OF PROPER NOUN

Detecting Proper nouns in English languages is not very difficult; Nouns name people, places, and things. Every noun can further be classified as common or proper. A proper noun has two distinctive features: it will name a specific (usually a one of a kind) item, and it will begin with a capital letter no matter where it occurs in a sentence. Detecting Proper noun is quite challenging in Arabic languages as it shares no cognates with English. The Arabic Information Retrieval proper name module utilizes clue words in the document text to detect Proper Names in six different categories: People , Major Cities , Locations , Countries , Organizations ,

Political parties and Terrorist Groups .

To detect proper nouns in Arabic text we use set of keywords to guide us to the place where we can find them in the text. By using keywords we mark name phrases that might contain a certain proper noun then we process these phrases to extract proper nouns. One way to process these phrases and extract the names is to construct a bunch of heuristic rules and use them to parse the phrase to extract the name. This technique has many limitations: it is hard to tell exactly where the name starts and where it ends in the phrase especially for foreign names, e.g. Bill Clinton . Each person writes in a different way with a different style, so the same name phrase can be written in many different ways, since no matter how many rules you add to the system you will never cover all the scenarios that you might face.

In this paper we described a new technique to process the phrases to extract the proper nouns by creating set of keywords to tag the proper noun in the text we look for the keywords and special verbs in the text to mark the proper noun, this keyword usually followed by a Proper Noun. The paper answers two major questions: where we can find names in the text and how to extract them.

We generated a set of rules to predict where the names are located in the text. These rules are based on two things: the keyword and some special verbs. Names seem to appear close to one of these keywords or special verbs in Arabic text. To mark the proper noun in the text we look for the keywords and special verbs in the text to mark the name phrases [20] we classified them in different classes: people, locations, organizations, events and products. Table1 shows some examples of these keywords and special verbs.

TABLE 1
KEYWORDS AND SPECIAL VERBS

KEYWORD /SPECIAL VERB		KEYWORD/ SPECIAL VERB
Mr.		Announced
President		Newspaper
Professor		Bank
Country		Sea
City		Mother
Conference		Father
Exhibit		Son
War		Republic
Said		Ministry

The location entity is recognized by the rule that stipulates: If we have in the text a word whose lemma is in this list ( ) followed by a Proper Noun, this sequence of words is marked as a location.

For example, in the Arabic text " ", one named entity is recognized as Location

We presented a prefix for personal names such as (Mr., Dr., Majesty, Sir, etc…), place names a prefix such as (city, country, republic, kingdom, etc…), in this system to retrieve names (surname, middle name, last name) we must write, " "or " " between two names.

6. EXTRACTING PROPER NOUN

Algorithm steps to extract proper noun in Arabic language is described as follows:
* Remove diacritics

- Diacritics: special marks are put above or below the characters to determine the correct pronunciation. Such as " " e.g.
to Arabic (language).
* Remove punctuation and non letters. Such as " . ! ".
* Search in keyword file and special verbs using set of rule.
* Check for the prefix and strips off "
"
and remove it .
* Check for suffixes, "
etc."
* Extract the words that follow keywords and save them in the proper noun database. The rule for word that always followed by proper noun:
- Any word follows any of the following words " "((must be a proper noun.
-Any word follows any of the following words " "((must be a proper noun
Such as:
- Any word follows any of the following words " " which means kunai must be a proper noun.
- Any word follows any of the following words "" must be proper noun.
- Any word follows any of the following words " " or " " must be noun.
- the combination of followed by often one of the Muslim 99 names of god such as must be a proper noun.
- Any word follows any of the following words Prepositions " :
" and has the pattern must be proper noun, such as in the following example :

Figure 5 below shows a flowchart to extract proper noun from Arabic language text.

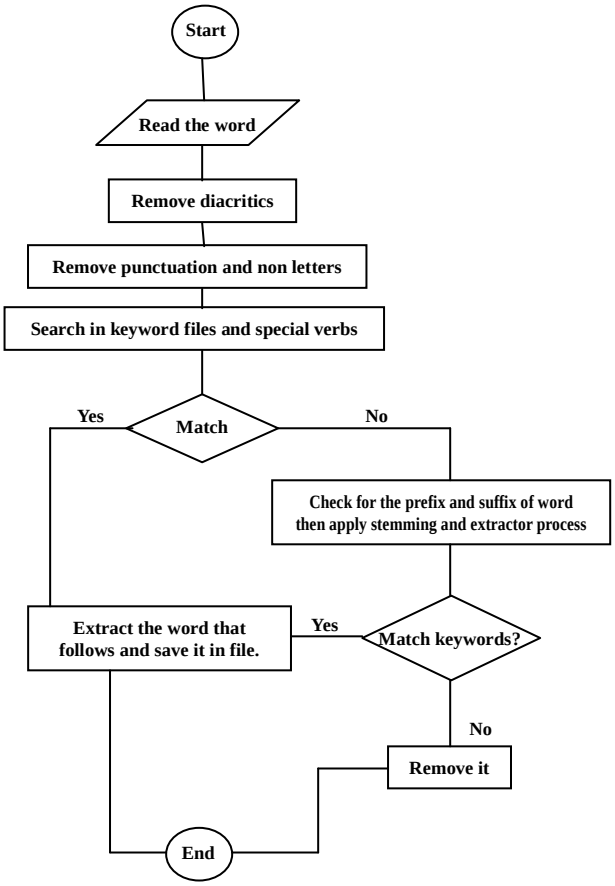

Figure 5: The automatic Algorithm for extracting proper noun

7. PERFORMANCE EVALUATION

We have evaluated our new technique to extract the proper noun using 20 randomly documents selected from the Al-Raya newspaper published in Qatar, and Alrai newspaper published in Jordan.

We classified the proper noun into 7 sub categories, table 2 below shows the categorizer's of proper noun and shows precession of the detection for each class.

We classified proper noun according to major class (Location, Organization, Person name, Equipment, Scientific, Temporal, and event) and sub class (city, company, ism, software, disease, date, conference, etc.), to compute Precession we use the following formula:
Precession =Total # correct / (Total # correct + Total # incorrect)

TABLE 2
EFFECTIVENESS OF DIFFERENT CATEGORIES OF PROPER NOUN

Category	**Total # correct**	**Total # incorrect**	**Precession**
Location	165	15	91.6%
Person name	90	21	81.1%
event	31	5	86.1%
Organization	27	9	75%
Temporal	17	2	89.4%
Equipment	11	3	78.5%
Scientific	7	1	87.5%
Total	348	56	86.1%

CONCLUSION

This paper proposed a new Arabic technique that enables to retrieve proper nouns in the Arabic text using Keywords. We generate a set of rules to state where the proper nouns are located in the text. These rules are based on two things: the keywords and some special verbs. To mark the proper noun in the text we look for this keywords and special verbs in the text and then apply rules to extract proper noun. We extract 86.1% of the proper noun found in the text. The difficulty of this work is how to extract proper nouns from text if it is not contain keywords. We plan to expand our method to include extract proper nouns using individual names, keywords, and root of the Arabic name.

REFERENCES

[1] Aljlayl, M. and Frieder, O. (2001), "Effective Arabic-English Cross-Language Information Retrieval via Machine Readable Dictionaries and Machine Translation", *ACM Tenth Conference on Information and Knowledge Management*, Atlanta, Georgia, November.

[2] Allan, J. and Raghavan, H. (2002), "Using part-of-speech patterns to reduce query ambiguity", *In Proceedings of SIGIR-02*, Tampere, Finland.

[3] Egyptian Demographic Center, 2000. http://www.frcu.eun.eg/www/homepage/cdc/cdc.htm

[4] Tayli, M., and Al-Salamah, A. (1990), "Building Bilingual Microcomputer Systems", In *Communications of the ACM*, Vol. 33, No.5, Pages 495-505.

[5] Al-Daimi, K., and Abdel-Amir, M. (1994), "The Syntactic Analysis of Arabic by Machine", *Computers and Humanities*, Vol. 28, No. 1, pp. 29-37.

[6] Imed Al-Sughaiyer i, and Ibrahim Al-Kharashi (2000). "An Efficient Arabic Morphological Analysis Technique for Information Retrieval Systems". In *ACIDCA'2000 International Conference*. Monastir, Tunisia, March.

[7] Al-Shalabi, R, and Kanaan, G. (2004), "Constructing An Automatic Lexicon for Arabic Language", International Journal Of Computing & Information Sciences, vol.2, no.2, August, page 114,128.

[8] Kenneth R. Beesley (1998), "Consonant Spreading of Arabic Stems", In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*.

[9] Freeman, A., and Condon, S. and Ackerman, C. (2006), "Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm", *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, June, pages 471–478,New York.

[10] Rau L. (1991). "Extracting Company Names from Text". *Proceedings of the Seventh Conference on Artificial Intelligence Applications*. Miami Beach, Florida.

[11] Imed Al-Sughaiyer i, and Ibrahim Al-Kharashi (2000). "An Efficient Arabic Morphological Analysis Technique for Information Retrieval Systems". In

ACIDCA'2000 International Conference. Monastir, Tunisia, March.

[12] Al-Shalabi, R, and Evens, M. 1998. "A Computational Morphology System for Arabic". *Workshop on Computational Approaches to Semitic Languages, COLING -ACL*.

[13] Al-Fedaghi, S., Al-Anzi, F. (1989), "A new algorithm to generate Arabic root-pattern forms.", *Proceedings of the 11th National Computer Conference*, King Fahd University of Petrolium & Minerals, Dahran, Saudi Arabia., pp04-07.

[14] Abuleil, S., and Evens, M. (1998), "Discovering Lexical Information by Tagging Arabic Newspaper Text", *Proceedings of the Workshop on Semitic Language Processing*. COLING-ACL'98, Aug. 16, pp. 1–7.

[15] Abuleil, S. and Alsamara, K., "New Technique to Support Arabic Noun Morphology: Arabic Noun Classifier System (ANCS)", *International Journal of Computer Processing of Oriental Languages*, Vol. 17, No. 2 (2004) 97–120

[16] Coates-Stephens, S. (1992), "The Analysis and Acquisition of Proper Names for Robust Text Understanding". Unpublished doctoral dissertation, City University, London.

[17] Grishman, R. (1997), "Information extraction: Techniques and challenges". *Summer Convention on Information Extraction (SCIE)*, 10–27.

[18] Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0 Linguistic Data Consortium (LDC) catalog number LDC2002L49 and ISBN 1-58563-257-0,.

[19] Abuleil, S. (2003), "Extracting Names from Arabic text for question-answering systems".

[20] Church, Kenneth (1988), "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", *Proceedings of Second Conference on Applied Natural Language Processing*, pp. 136-143.

[21] http://en.wikipedia.org/wiki/Arabic_name

" " . [23]

" "

1996. [24]

" "

1989. [25]

" "

1987.[26]

1999. [27]