Building a Proper Noun Database to Support Natural Language Applications

Saleem Abuleil **Martha Evens**
Computer Science Department, Illinois Institute of Technology
10 West 31 Street, Chicago IL 60616
abulsal@charlie.cns.iit.edu **mwe@math.nwu.edu**

Abstract

In this paper we describe a system for building an Arabic proper noun database by parsing Arabic newspaper text. We are using several algorithms, techniques and rules for marking each proper noun in the text, extracting the information about it and adding it to our lexical database.

Introduction

The lexicon is the backbone of any natural language application. It is an essential basis for parsing, text generation, question answering, and information retrieval systems. The best way to find the necessary information, we believe, is to extract it automatically from the text. We are building a proper noun database automatically, which requires us to solve the problem of identifying the proper nouns in the Arabic text?

The Arabic language does not distinguish between lower and upper case letters. Upper case letters provide major help in marking the proper noun in English; they allow us to look for the capitalized letters in the text and start working from them. In Arabic there is no clear rule like this to guide us to find proper nouns, which leaves us with a big problem in recognizing them in Arabic text. In this paper we are trying to find the answer to this challenge through building a system that parses an Arabic text, marks the proper nouns in it, and extracts all information about them for insertion in the lexicon. We have used a corpus from the Qatar newspaper Al-Raya.

Background

Constructing lexical entries for proper nouns is not less important than defining and analyzing common nouns, verbs, and adjectives for supporting natural language applications. [Mehdi 1986] describes a computer system for syntactic parsing of the Arabic sentences. The system is implemented by using Definite Clause Grammar (DCG) formalism and prolog has been used as a programming language. [Ibrahim, Clarke, and Fahmy 1989] have suggested a framework to deal with the affixational aspect of the Arabic language. [Foxley and Feddag 1990] adopted a strategy combined affixes to alleviate the speed of operation overhead that the affix manipulation routines can take. [Feddag and Foxley 1991] provided a single powerful framework for intelligent database where the system stores only roots of the verbs and uses a program intelligent enough to automatically handle all derived forms. The semantic categories of proper nouns are crucial information for text understanding [Wolinski et al. 1995] and information extraction [Cowie and Lehnert 1996]. They are also used in information retrieval systems [Paik et al. 1993]. Rau [1991] argues that proper nouns not only account for a large percentage of the unknown words in a text, but are also recognized as a crucial source of information in a text for extracting contents, identifying a topic in a text, or detecting relevant documents in information retrieval. Wacholder [1997] analyzed the types of ambiguity - structural and semantic - that make the discovery of proper names in the text difficult. Jong-Sun Kim and Evens [1995] built a natural language processing system for extracting personal names and other proper nouns from the *Wall Street Journal*.

Parsing Proper Noun

Each type of proper noun has different attributes, so we setup separate tables for each.

Personal names:

proper noun	**occupation**	**organizat -ion**	**nationality**
Evens Martha	Professor	IIT	American

Organization names:

proper noun	**type**	**location**	**service**
IIT	university	Chicago	education
Byte	magazine	America	computer

Location (political names):

proper noun	type	location	language
Chicago	city	Illinois	English
Illinois	state	America	English
USA	country	World	English

Location (natural geographical names):

proper noun	type	location
Nile	river	Africa
Atlantic	ocean	world

Times:

proper noun	part-of	located-at
September	months	9th
Christmas	holidays	December

Products:

product name	kind-of	made-in
Toyota	vehicle	Japan
Compaq	computer	America

Events:

event-name	type	place	year	special-ist-on
Al-Kitab	exhibition	Egypt	1995	books
Madrid	conference	Aspen	1993	peace

Category (nationality, language, religion, ethnic, party, etc.):

proper noun	type	related-to
American	nationality	America
Arabic	language	Arabs

The Arabic language does not distinguish between upper/lower case letters like the English language. This makes it not nearly as easy to locate proper nouns in Arabic text as in English text. For this reason we will use another technique for tagging the proper nouns in the text. This technique depends on the keywords. We have studied, analyzed, and classified these keywords, to use them to guide us in tagging the proper nouns in the text and figuring out the types and the features. We have classified these keywords as follows:

- Personal names (title): **Mr.** John Adams
- Personal names (job title): **President** John Adams
- organization names: Northwestern **University**
- Locations (political names): **State** of Illinois
- Location (natural names): **Lake** Michigan
- Times: **Month** of September
- Products: IBM **Computer**
- Events: **Exhibition** of Egyptian books
- Category: Arabic **Language**

We have also developed a set of grammatical rules to parse the proper noun phrases in the text:

LPNP → SPNP K/W SPNP
| K/W SPNP SPNP
| SPECAIL-VERB SPNP
| SPNP
SPNP → K/W-PERSON LPN-OCCUPATION
| ال K/W-PERSON LPN-TITLE
| ال K/W-ORG-LOC LPN-ORG-LOC-1
| K/W-ORG-LOC LPN-ORG-LOC-2
| K/W-PRODUCT LPN-PRODUCT
| ال K/W-PRODUCT LPN-PRODUCT
| K/W-TIME LPN-TIME
| K/W-EVENT LPN-EVENT
| ال K/W-EVENT LPN-EVENT
| K/W-CATEGORIES LPN-CATEGORIES
| K/W-CATEGORIES LPN-CATEGORIES
LPN-OCCUPATION → K/W-ORG/LOC PN-ORG/LOC PN-PERSON
| K/W-ORG/LOC PN-ORG/LOC
| PN-ORG/LOC PN-PERSON
| PN-ORG/LOC
| ال K/W-ORG/LOC ال PN-ORG/LOC PN-PERSON
| ال K/W-ORG/LOC ال PN-ORG/LOC
LPN-TITLE → PN-PERSON
| ADJECTIVE
| ADJECTIVE PN-PERSON
| PN-PERSON ADJECTIVE
| ADFPN
| ADFPN PN-PERSON
| ADFPN ADJECTIVE
| ADFPN ADJECTIVE PN-PERSON
| ADFPN PN_PERSON ADJECTIVE
LPN-ORG-LOC-1 → ADJECTIVE
| ADJECTIVE PN-LOC-ORG
| ADFPN
| ADFPN PN-ORG-LOC
| ADFPN ADJECTIVE
| ADFPN ADJECTIVE PN-ORG-LOC
LPN-ORG-LOC-2 → PN-ORG-LOC
| PN-ORG-LOC ADJ
| PN-ORG-LOC ADFPN
| PN-ORG-LOC ADFPN ADJ
| PN-ORG-LOC ADJ ADFPN
| PN-ORG-LOC PN-PERSON
| PN-ORG-LOC ال ADJ PN-PERSON
| PN-ORG-LOC ال ADFPN PN-PERSON
LPN-PRODUCT → PN-PRODUCT
| ADFPN | ADJ
| PN-PRODUCT ADFPN

| PN-PRODUCT ADJ
| PN-PRODUCT ADFPN ADJ
| ADJ PN-PRODUCT
| ADFPN PN-PRODUCT

LPN-TIME → PN-TIME
| PN-TIME الموافق PN-TIME

LPN-EVENT → PN-EVENT1 NP
| PN-EVENT1

LPN-EVENT1 → PN-EVENT
| ADJ | ADFPN
| PN-EVENT ADJ
| ADFPN ADJ
| ADFPN PN-EVENT

NP → لل NOUN

LPN-CATEGORIES → PN-CATEGORIES
| ADFPN

ADFPN → ال (ADJECTIVE DERIVED FROM PROPER NOUN)

The (ال) is the determiner in the Arabic language, it tells whether the proper nouns come right next to the keyword or not.

Parser System

Our system consists of three main subsystems beside the lexical database as shown in Figure 1.

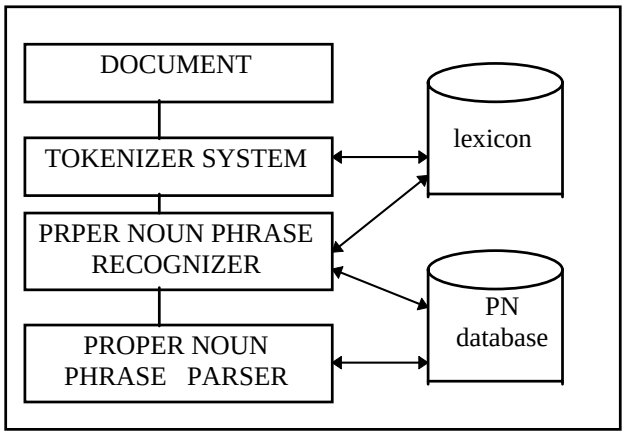


Figure 1: Organization of the Parser

Tokenizer System

We have implemented an algorithm that can isolate the punctuation marks as well as isolate the extra particles attached to the beginning of the word, while they are not part of it. We have classified the words in the Arabic language into eight categories with respect to their prefixes. This system caries out three main steps: Isolate the word from the text, pass it to a certain algorithm to classify it, and with respect to this classification run a certain algorithm to generate the token.

Recognizer for Proper Noun Phrases

From our analysis of *Al-Raya* newspaper text we have categorized the proper nouns with respect to the way they occur in the text into three categories: proper nouns that start with a keyword, proper nouns that end with a keyword and proper nouns without an adjacent keyword. We have implemented this system for proper noun phrases that start with a keyword, end up with one of the following: verb, particle, pronoun, auxiliary verb or punctuation mark, and include one or more of the following: keywords, proper noun, adjective, adjective derived from proper noun, word of unknown type. Figure 2 shows how this system works.

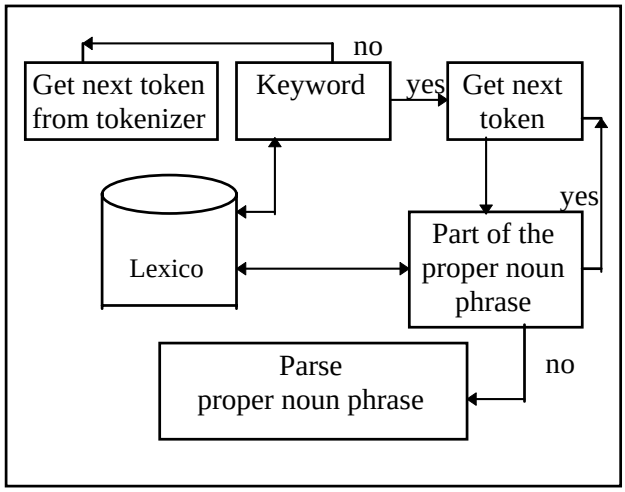


Figure 2: Recognizer for Proper Noun Phrases

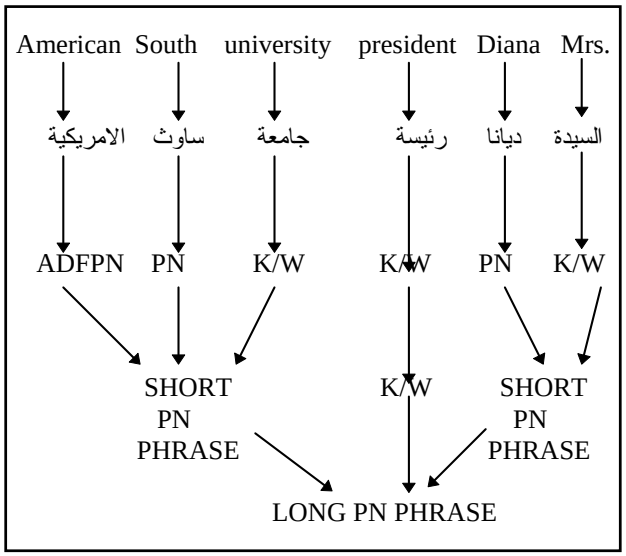
Parser for Proper Noun Phrase

This system is responsible for parsing proper noun phrases to mark the proper nouns, find their features, and extract further information. This system uses two techniques to accomplish this task, a set of grammatical rules to mark the proper noun and to extract the information about it and the keyword to find the type and features.

Example:

قامت السيدة ديانا رئيسة جامعة ساوث الامريكية بزيارة ال معرض شيكاغو الكبير للكمبيوتر

Mrs. Diana president of American's South University made a visit to the big computer exhibition in Chicago



PN: proper noun. **K/W**: keyword
ADFPN: adjective derived from PN

Databases

We are using two kinds of databases to assist the system. The first database contains the words, their types, and their features. This is the lexical database that developed by Abuleil and Evens [1998]. The second database is the one that we are building for the proper nouns. It contains main table with all the proper nouns, the proper noun itself, the type (person, organization, location, group, time, event) and the gender (male or female), and there are additional tables connecting to the main table each one with information about one type of proper noun as shown in Figure 3.

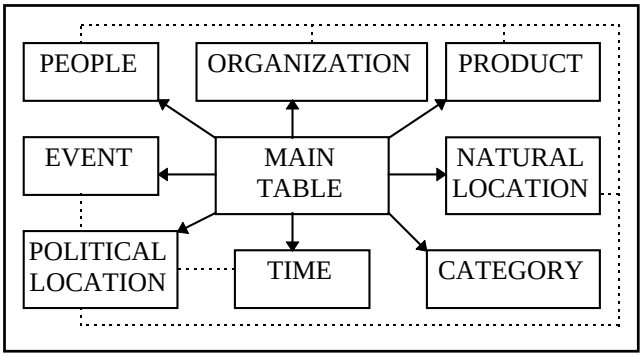


Figure 3: The Proper Noun Tables

Results

We have tested the system on 40 articles from a corpus developed by Ahmad Hasnah[1996] based on text given to Illinois Institute of Technology, by the newspaper, *Al-Raya*, published in Qatar. We have got the following result: the actual proper nouns in the text (isolated proper noun or proper noun with adjacent keyword) are 175, the first part of the system (proper noun phrases recognizer) which is responsible for marking the proper noun phrases, found 152 - 87% of the proper noun phrases (see Figure 4). The second part of the system (proper noun phrases parser) which is responsible for parsing the proper noun phrases, found 147 proper nouns, this amounted to 96% of the proper noun phrases that are found by the first system (proper noun phrases recognizer) see Figure 5, and 84% of the total number of proper nouns in the text. In summary total the parser system by using both the proper noun phrase parser and the accumulated proper noun database, the one that we are building, found and marked 161 or 92% of the total number of proper nouns in the text (see Table 1).

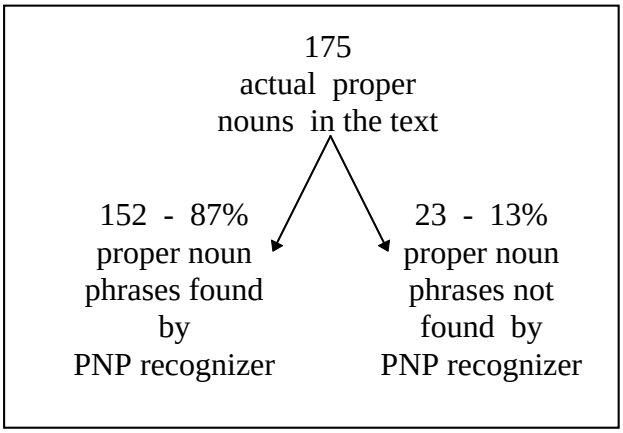


Figure 4: Proper Noun Phrase Recognizer System

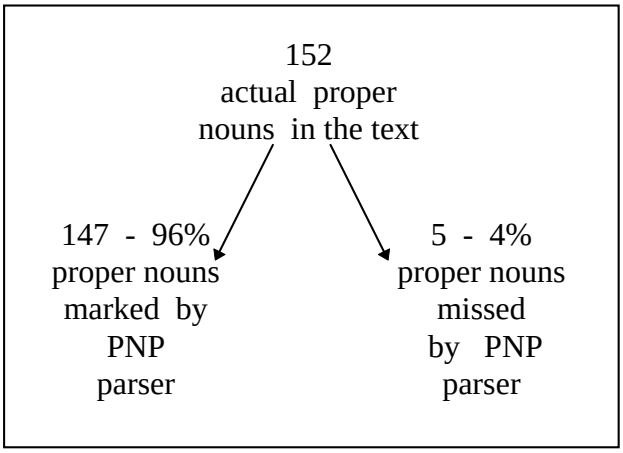


Figure 5: Proper Noun Phrase Parser System

PN in the text	PN marked by Parser System
175	161 - 92%

Table 1: Proper Nouns Marked by the Parser System

our parser also produced a few garbage nouns, that the parser marked as proper nouns though they are not. The problem is that in some cases the parser could not distinguish between proper noun and some other special adjectives and nouns. We believe that we will reduce this problem as we parse more text and identify the special words that create this difficulty.

Conclusion

In order to build a large database of Arabic proper nouns automatically we have classified the proper nouns in the Arabic language into different categories and used a new technique to tag them from the Arabic text by using keywords. We developed a set of grammatical rules for this purpose. We have developed a system using these rules that extracts the proper nouns and the information about them from newspaper text.

References

Abuleil, S. and Evens, M., 1998. Discovering Lexical Information by Tagging Arabic Newspaper Text. *COLING-ACL '98*, 1-7. University of Montreal, Quebec, Canada, Aug 16 1998.

Cowie, J., and Lehnert, W., 1996. Information Extraction *Comm, of the ACM* 39(1):83-92.

Elmi. M. A. 1994. A Natural Language Parser with Interleaved Spelling Correction Supporting Lexical Functional Grammar and Ill-Formed Input, Unpublished Ph.D. Dissertation, Computer Science Department, Illinois Institute of Technology, Chicago IL.

Feddag, A., Foxley, E., 1991. An Intelligent Lexical Analyzer for Arabic and English. 13th Research Conference on Information Retrieval. British Computer Cociety (BCS). Lancaster University, 8-9th April, UK.

Foxley, E., Feddag, A., 1990. A Syntactic and Morphological Analyzer of Arabic Words. *Proceedings of the Second International Conference on Computing in Arabic-English*. Cambridge University, UK.

Hasnah, A., 1996. Full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval For Arabic Documents. Ph.D. Dissertation, Illinois Institute of Technology, Chicago, IL.

Ibrahim, A., Douglas, J., Fahmy, A., 1989. Arabic Machine Translation. *Proceedings of the First International Conference in Computing in Arabic-English.* Cambridge University, UK.

Kim, J-S., and Evens, M., 1995. Extracting Personal Names from the Wall Street Journal. *Proceedings of the 6th Midwest Artificial Intelligence and Cognitive Science Society Conference*, 78-82. Carbondale, IL, April 21-23.

Mehdi, S. A. 1986. Arabic Language Parser, International Journal of Man-Machine Studies. 25(5):593-611.

Paik, W., Liddy, E. D., Yu, E., and Mckenna, M., 1993. Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. In B. Boguraev and J. Pustejovsky 44-54. eds, *Corpus Processing for Lexical Acquisition*, Cambridge, Mass.: MIT Press.

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

Wacholder, N., Ravin, Y., and Choi, M., 1997. Disambiguation of Proper Names in Text. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 202-208. Mar 31- Apr 3, Washington, DC.

Wolinski, F., Vichet, F., and Dillet, B., 1995. Automatic Processing of Proper Names in Text. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 23-30. Dublin, Ireland.