Discovering Lexical Information by Tagging Arabic Newspaper Text

Saleem Abuleil **Martha Evens**
CSAM, 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 lexicon automatically by tagging Arabic newspaper text. In this system we are using several techniques for tagging the words in the text and figuring out their types and their features. The major techniques that we are using are: finding phrases, analyzing the affixes of the words, and analyzing their patterns. Proper nouns are particularly difficult to identify in the Arabic language; we describe techniques for isolating them.

INTRODUCTION

A lexicon is considered to be the backbone of any natural language application. It is an essential basis for parsing, text generation, and information retrieval systems. We cannot implement any of these applications or others in the natural language area without having a good lexicon. All natural language processing systems need a lexicon full of explicit information [Ahlswede and Evens, 1988; Byrd et al., 1987 McCawley, 1986]. The best way to find the necessary lexical information, we believe, is to extract it automatically from text.

We are developing a part-of-speech tagger for Arabic newspaper text. We are testing it on a corpus developed by Ahmad Hasnah [1996] based on text given to Illinois Institute of Technology, by the newspaper, Al-Raya, published in Qatar. The questions we address here are how to build an efficient techniques for automating the tagger system, what techniques and algorithms can be used in finding the part of speech and extracting the features of the word. When it comes to the Arabic language there are problems and challenges that are not present in English or other European languages.

- Newspaper articles are full of proper nouns that need special rules to tag them in the text, because the Arabic language does not distinguish between lower and upper case letters, which leave us with a big problem in recognizing proper nouns in Arabic text.

- The lack of vowels in the text we are using creates big problems of ambiguity. Different vowels change the word from noun to verb and from one type of noun to another; they also change the meaning of the word. For example, the following two words have the same letters with the same sequence but with different vowels. The result is different meanings.

  كتب k(a)t(a)b wrote

  كتب k(u)t(u)b books

  Most published Arabic text is not vowelized with the exception of the Holy Quran and books for children.

- Some words in Arabic text begin with one, two, three, or four extra letters that constitute articles or prepositions. For example, the following word consists of two parts: the particle (a preposition letter) that is attached to the beginning of the noun while it is not part of it and the noun itself.

  (on occasion) بمناسبة ➔ ب + مناسبة

We need to identify these cases in the text and deal with them in a perceptive way.

In this paper we are trying to find answers to these challenges through building a tagger system whose main function is to parse an Arabic text, tag the parts of speech, and find out their features to build a lexicon for this language. Three main techniques used in this system for tagging the words are: finding phrases (verb phrases, noun phrases, and proper noun phrases), analyzing the affixes of the word, and analyzing its pattern.

1. TAGGING VERB AND NOUN

There are several signs in the Arabic language that indicate whether the word is a noun or a verb. One of them is the affix of the word: some of the affixes are used with verbs; some of them are used with nouns; and some of them are used with verbs and nouns. A lot of research projects have used this technique to find the part of speech of a word. Andrei Mikheev [1997] used a technique for fully automatic acquisition of rules that guess possible part-of-speech tags for unknown words using their starting and ending segments. Several types of guessing rules are included: prefix morphological rules and suffix morphological rules. Zhang and Kim [1990] developed a system for automated learning of morphological word function rules. This system divided a string into three regions and inferred from training examples their correspondence to underlying morphological features. More advanced word-guessing methods use word features such as leading and trailing word segments to determine possible tags for unknown words. Such methods can achieve better performance, reaching a tagging accuracy of up to 85% on unknown words for English [Brill 1992; Weischedel et al., 1993]. Another sign that indicates whether a word is a noun or a verb is the pattern. In the Arabic language the patterns function as an important guide in recognizing the type of the word; some of these patterns are used just for nouns; some of them are used just for verbs; and others are used for both nouns and verbs. One more sign comes from grammatical rules; several grammatical rules can be used to distinguish between nouns and verbs, some letters in the Arabic language (letters of signification are similar to prepositions in the English language) mark the nouns; others mark the verbs

2. TAGGING PROPER NOUNS

Constructing lexical entries for proper nouns is not less important than defining and analyzing common nouns, verbs, and adjectives for supporting natural language applications. 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]. A number of studies have shown the usefulness of lexical-semantic relationships in information retrieval systems [Evens et al., 1985; Nutter et al., 1990; Abu-Salem, 1992]. The lexical-semantic relationships are also important in other applications like question-answering systems [Evens and Smith, 1978]. Rau [1991] argues that proper nouns not only account for a large percentage of the unknown words in a text, but also are 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.

We have classified the proper nouns that we found in the Al-Raya newspaper as follows:

Personal names:

proper noun	occupation	organization	nationality
M. Evens	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

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.):

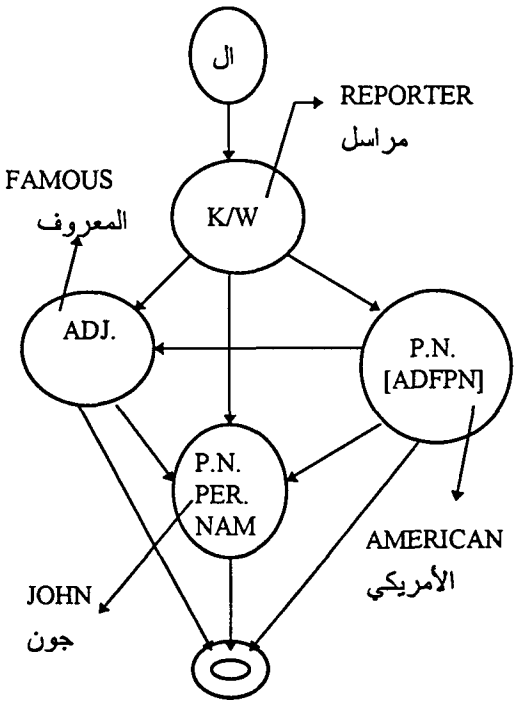
proper noun	type	related-to
American	nationality	America
Arabic	language	Arabs

The Arabic language does not distinguish between upper/lower case letters like the English language. So the proper nouns do not begin with a capital letter. This makes it not nearly as easy to locate them 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
- 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 identify the proper noun phrases in the text. Example:

PNP → ال K/W-TITLE A
A → A1 | A2
A1 → ADFPN
| ADFPN PN-PERSON
| ADFPN A2
A2 → ADJ
| ADJ PN-PERSON
ADFPN → ال [ADJ. DERIVED FROM P.N.]



3. TAGGER SYSTEM

This system consists of four main subsystems beside the database (lexicon):

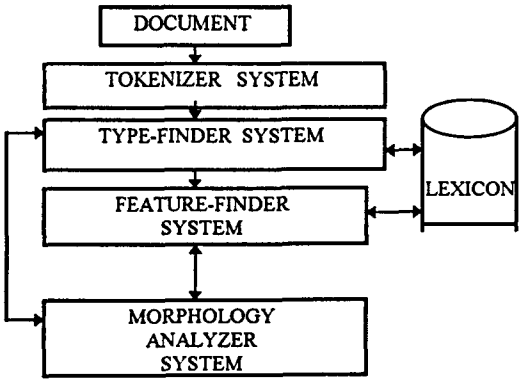
Tokenizer System: This system locates a document and isolates the words (tokens).

Type-Finder System: The main function of this system is to get the token from the tokenizer system, to get some information about it from the morphology analyzer system, to go through several tests one by one until we find the part of speech of the word.

Feature-Finder System: This system is responsible for finding the features of the word (gender, number, person, tense). It sends the word to the morphology analyzer system, gets back information about it, analyzes this information, and figures out the features of the word.

Morphology analyzer system: This system is used by both the type-finder system and the feature-finder system to analyze the suffix and prefix of the word. This system contains three subprograms: one for nouns, one for verbs, and one for particles. The main function of these algorithms is to isolate the affixes of the word and find the gender, number, person, and tense.

Database (the lexicon): We started from a hand built lexicon created by Khalid Alsamara [1996], which our system uses and constantly updates. The lexicon consists of the main table and several tables connected to it one for verbs, one for nouns, one for particles. We add several tables for proper nouns.



4. 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 prefix. 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 we run a certain algorithm to generate the token.

5. TYPE-FINDER SYSTEM

This system goes through several tests starting by checking the database, identifying the phrases, analyzing the affixes of the word, and analyzing its pattern.

5.1 PHRASE-FINDER TEST

After we check the database and discover that our token is absent, we move to the second test. The phrase-finder test uses a set of grammatical rules that identify the phrases in the text. It looks for the phrases (verb phrase, noun phrase, and proper noun phrase) in the text, analyzes them and figures out the part of speech of the word.

Example:

قامت السيدة ديانا بزيارة إلى مؤتمر الحاسب الذي يعقد للمرة الأولى في مدينة شيكاغو

Mrs. Diana made her visit to the computer conference, that is being held for the first time in Chicago

This test determines the part of speech of all the underlined words.

NP → PART-N N
NP → PART-1 N
NP → الى N
NP → الى مؤتمر

so مؤتمر (conference) is a noun

PNP → K/W[T:x, G:y] P.N[T:x, G:y]
PNP → السيدة [T:person, G:feminine]
P.N[T:person, G:feminine]
PNP → السيدة [T:person, G:feminine]
ديانا [T:person, G:feminine]

so ديانا (Diana) is a proper noun, for a female human being.

5.2 CHECKING THE AFFIX PATTERNS

If the second test fails to identify the part of speech of our token, we continue to the third and the four test in sequence, for these two tests we are using two techniques: analyzing the affixes of the word and finding its patterns.

First, we have classified affixes into two groups:

- affix rule (A): if an affix occurs in a word we can surely determine the type of the word without any doubt.

  Example:

Prefix	Suffix	Type
---	ات | ة	NOUN
---	تن | وا	VERB

- Affix rule (B): if an affix occurs in a word we can usually ascertain the type of the word.

Example:

Prefix	Suffix	Type
ي | سي	ان | ون	verb
ا	ى	noun

Second, we have collected one hundred and sixty three patterns that cover the patterns in the Arabic language, we have classified these patterns according to the type of the word they are used for (noun, verb, or noun and verb).

Example:

انفعل (anf9l) فعلل (f9ll) افتعل (aft9l)
used with **verb**
فعال (f9al) مفعله (mf9lh) فعاله (f9alh)
used with **noun**
فعل (f9l) فاعل (fa9l) تفعل (tf9l)
used with **noun** and **verb**

5.2.1 AFFIX-RULE-(A) TEST

This third test gets the affix of the word from the morphology analyzer system and checks these affixes with affix rule (A). If there is a match then the test succeeds otherwise we continue to the fourth test.

Example:

قامت السيدة ديانا بزيارة إلى مؤتمر الحاسب الذي يعقد للمرة الأولى في مدينة شيكاغو

Mrs. Diana made her visit to the computer conference, that is being held for the first time in Chicago.

This test determines the part of speech of all the underlined words.

word	prefix	suffix	result
السيدة	ال [e.l]	ة	noun
بزيارة	ب [e.l]	ة	noun

5.2.2 PATTERN-AFFIX-RULE-(B) TEST

The fourth test uses a combination of affix rule (B) and the patterns of the word. This test uses affix rule (B) to support the decision that will be taken from the pattern technique. It gets the affix of the word from the morphology analyzer

system, checks the affix with affix rule (B) to find out if there is a match, finds the pattern of the word, analyzes it to find out the type that it is used for. We then go through the following table to get the final result for this test.

Type of the token from PATTERNS	Type of the token from AFFIX RULE (B)	RESULT
NOUN	X	NOUN
VERB	X	VERB
NOUN/VERB	X	X
FAIL	X	FAIL

So if we have a certain token, its pattern shows it is a NOUN or VERB, and its affixes show it is a NOUN with respect to affix rule (B), then with respect to our table this token should be a NOUN.

Example:

قامت السيدة ديانا بزيارة إلى مؤتمر الحاسب الذي يعقد للمرة الأولى في مدينة شيكاغو

Mrs. Diana made her visit to the computer conference, that is being held for the first time in Chicago.

This test determines the part of speech of all the underlined words.

RESULT FROM Affix rule (b)				RESULT FROM Pattern			R
W	S	P	R	T	PT	R	
قامت	ت	–	v	قام	f9l	v / n	v

W: word, S: suffix, P: prefix,
R: result T: token, PT: pattern

6. FEATURE-FINDER SYSTEM

This system sends the word and its type to the morphology analyzer system, gets back morphological information about it (the affixes and the gender, number, person, and tense of the word), gets the pattern of the word, analyzes this information using a certain rules we have developed for this system, and finds the features.

Example (1):

word	type	information from morphology Analyzer system	result interpreting the information
تقوم	verb	tense-4 / agent-14	pres / masc, 2nd, sing pres / femi, 3rd, sing
مؤتمر	noun	noun-2	masc, sing

Example (2):

word	pattern	used for
جبال	f9al فعال	plural / masculine
صحاري	f9ale فعالي	plural / feminine

8. CONCLUSION

We badly need a large integrated comprehensive lexicon. To achieve this goal we need to build this lexicon automatically. To build such a lexicon we are developing a part of speech tagger for Arabic text that extracts features of the words encountered. We have described three major techniques that we are using in this paper: finding phrases, analyzing the affixes of the word, and analyzing its patterns. We have classified the proper nouns in the Arabic language to different categories, we used a new technique to tag them from the Arabic text by using the keywords. We developed a set of grammatical rules for this reason.

REFERENCES

Abu-Salem, H., 1992, A Microcomputer Based Arabic Bibliography Information Retrieval System with Relational Thesauri (Arabic-IRS).

Unpublished Ph.D. Dissertation, Computer Science Department, Illinois Institute of Technology, Chicago, IL.

Ahlswede, T.E., and Evens, M., 1988, "Generating a Relational Lexicon from a Machine Readable Dictionary", International Journal of Lexicography, Vol. 1, No. 3, pp. 214-237.

Alsamara, K., 1996. An Arabic Lexicon To Support Information Retrieval, Parsing, and Text Generation. Unpublished Ph.D. Dissertation, Illinois Institute of Technology, Chicago, IL.

Brill, E., 1992. "A Simple Rule-based Part of Speech Tagger". Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, pp. 152-155.

Cowie, J., and Lehnert, W., 1996. "Information Extraction", Communications of the ACM, Vol. 39, No. 1, pp. 83-92.

Evens, M., and Smith, R., 1978. "A Lexicon for Computer Question Answering System". American Journal of Computational Linguistics, Microfiches 81, pp. 16-24, and 83, pp. 1-98.

Evens, M., Vandendorpe, J., and Wang Y., 1985, "Lexical-Semantic Relations in Information Retrieval". In Humans and Machines. S. Williams, (ed.), Ablex, Norwood, NJ, pp.73-100.

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.

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, Carbondale, IL, April 21-23, pp. 78-82.

McCawley, J. 1986. "What Linguists Might Contribute to Dictionary Making If They Could Get Their Act Together". In P. Bjarkman and V. Raskin (eds.), The Real World Linguistics, Ablex, Norwood, NJ, 1986.

Mikheev A., 1997. "Automatic Rule Induction for Unknown-Word Guessing". Computational Linguistics, Vol.23, No.3, September 1997. pp. 405-423.

Nutter, J. T., Fox, E., and Evens, M., 1990, "Building a Lexicon from Machine-Readable Dictionaries for Improved Information Retrieval", Literary and Linguistic Computing, Vol. 2, No. 5, pp.1-18.

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, eds, Corpus Processing for Lexical Acquisition, MIT Press, Cambridge, Mass, pp.44-54.

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

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

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., and Palmucci, J., 1993. "Coping with Ambiguity and Unknown Words through Probabilistic Models". Computational Linguistics, Vol.19, No.2, pp.359-382.

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, Dublin, Ireland, pp. 23-30.

Zhang, B.-T. and Kim, Y.-T., 1990. "Morphological Analysis and Synthesis by Automated Discovery and Acquisition of Linguistic Rules". Proceedings of the 13th International Conference on Computational Linguistics (COLING-90), pp. 431-435.