Classify Arabic Nouns for Morphology Systems

Saleem Abuleil **Martha Evens**

Information System Department, Chicago State University
9501 S. King Drive, Chicago, IL 60628
sabuleil@hotmail.com

Computer Science Department, Illinois Institute of Technology
10 West 31 Street, Chicago IL 60616
evens@iit.edu

Abstract

Noun morphology, verb morphology, lexicon and thesaurus represent essential resources for natural language processing applications. Many successful attempts have been discussed and implemented in Arabic verb morphology but very few have discussed Arabic noun morphology particularly for nouns that are not derived from verbs. The main challenge about Nouns in Arabic language, there is no clear rule to define the morphological information and generate the morphology features for them. Instead each group of nouns follows its own pattern. In this paper we describe a classifier system using a rule base that uses suffix analysis as well as pattern analysis to help the user to classify the nouns and identify the group that it belongs to support morphology systems.

Keywords: Noun Morphology, Pattern, and Classification

1 Introduction

A morphology system is the backbone of a natural language processing system. No application in this field can survive without a good morphology system to support it. The Arabic language has its own features that are not found in other languages. That is why many researchers have worked in this area. (Al-Shalabi, 1998) developed a system that removes the longest possible prefix from the word where the three letters of the root must lie somewhere in the first four or five characters of the remainder. Then he generates some combinations and checks each one of them against all the roots in the file.

(Anne de Roeck and Waleed Al-Fares, 2000) developed a clustering algorithm for Arabic words sharing the same verbal root. They used root-based clusters to substitute for dictionaries in indexing for information retrieval. (Beesley and Karttunen, 2000) described a new technique for constructing finite-state transducers that involves reapplying a regular-expression compiler to its own output. They implemented the system in an algorithm called compile-replace. This technique has proved useful for handling non-concatenative phenomena, and they demonstrate it on Malay full-stem reduplication and Arabic stem inter-digitations.

Most verbs in the Arabic language follow clear rules that define their morphology and generate their features. Those nouns that are not derived from roots do not seem to follow a similar set of well-defined rules. Instead there are groups showing family resemblances, rather like strong verbs in English. We believe that further work in Arabic computational linguistics requires the development of a pattern bank for nouns. This paper describes the tool that we have built for this purpose.

2 Nouns in the Arabic Language

A noun in Arabic is a word that indicates a meaning by itself without being connected with the notion of time. There are two main kinds of noun: variable and invariable. Variable nouns have different forms for the singular, the dual, the plural, the masculine, the feminine, the diminutive, and the relative. The diminutive form formed by adding the letter (ي) after the second letter of a declined noun, in order to indicate paucity, contempt or affection, (e.g., ثعيلب).

Variable nouns are again divided into two kinds: inert and derived. The inert noun is not derived from another word, i.e., it does not refer to a verbal root. Inert nouns are divided into two kinds: concrete nouns (e.g., lion), and abstract nouns (e.g., love). Derived nouns are taken from another word (usually a verb) (e.g. office); they have a root to refer to. A derived noun is usually close to its root in meaning. It indicates, besides the meaning, the concrete thing that caused its formation (in the case of the agent-noun), or underwent its action (in the case of the patient-noun), or any other notions of time, place, or instrument.

3 Noun Classification

In this paper we focus on the following nouns: substantive nouns, agent nouns, instrument nouns, adjectives, relative adjective, proper adjectives (adjectives derived from proper nouns), and adverbs. Some of these nouns are not derived from verbs and some are. All of these nouns use the same pattern when it comes to the dual form either for masculine or feminine, but there are many ways to form the plural noun. Some of the nouns have both masculine and feminine forms, some of them have just feminine forms and some have just masculine forms. A few nouns use the same format for both the plural and the dual (e.g. مدرسين teachers used for both dual and plural) For most nouns, when they end with the letter (ة), this indicates the feminine form of the noun, sometimes it does not, but it changes the meaning of the noun completely (e.g. مكتب office, مكتبة library). Sometimes the same consonant string with different vowels has different meanings (e.g. مدرسة school, مدرسة teacher). Nouns are not like verbs in the Arabic language, there is no clear rule to define the morphological information and generate the morphology features for them. Instead each group of nouns follows its own pattern.

We have classified the nouns into 152 groups according to their patterns for singular, plural, masculine and feminine. We generated a method for each group to be used to find the morphological information and to form its paradigm. Very few of these groups have a unique pattern for plural and singular; and most of them share the same pattern with other groups. Table 1 shows some examples of these groups and their patterns. The digit 9 stands for the letter “ayn [ع]”, ‘ stands for “hamzeh [ء]”, w stands for “waw [و]” and @ stands for “ta [ة]” since there are no corresponding letters in English for these letters.

Table 1. Pattern Classification

Group #	Rule				Ex.
	S-M	S-F	P-M	P-F	
1	F9L	X	X	aF9aL	قلم "klm"
8	F9l	X	X	F9aL	جمل "jml"
2	F9LwL	X	X	F9aLeL	جمهور "jmhwr"
3	X	F9L@	X	F9L	صورة "swr@"
4	Fa9L	Fa9L@	F9L@ Fa9Len/ Fa9Lwn	Fa9Lat	قاتل "qatl"
5	Fa9L	Fa9L@	F9aL Fa9Len/ Fa9Lwn	Fa9Lat	جامع "ja'9"
6	X	F9L@	X	F9al	قلعة "qlu@"
11	X	F9L@	X	aF9aL / F9L	شجرة "shjr@"
7	X	F9LLa'	X	F9aLL	سلحفاء "sl7fa'"
9	X	F9wL@	X	F9wLat	حكومة "7kwm@"

S: Singular F: Feminine P: Plural M: Masculine X: not available

All proper adjective nouns like أمريكي American and most relative adjective like رياضي athletic follow one of two rules: If the noun ends with (ا or ة) then drop it and add the suffix. If the noun does not end with (ا or ة) then just add the suffix. Very few relative adjective that fail to follow these rules like the noun "عرب" "Arab". Most of the paradigm follows the same rules but the masculine plural does not. The masculine plural for عربي is عرب and not عربيين or عربيون.

4 Noun Morphology System

The system reads the next noun in the text, isolates and analyzes the suffixes of the noun, generates its pattern, and uses either the DB Checker Module and Classified Noun Table, the Suffix/Pattern Analysis or the Noun Classifier Module to find the group to which the noun belongs, to identify the rules that apply to this group, to generate all morphological features with respect to the number and gender, and to update the database. The system consists of several modules as shown in Figure 1.

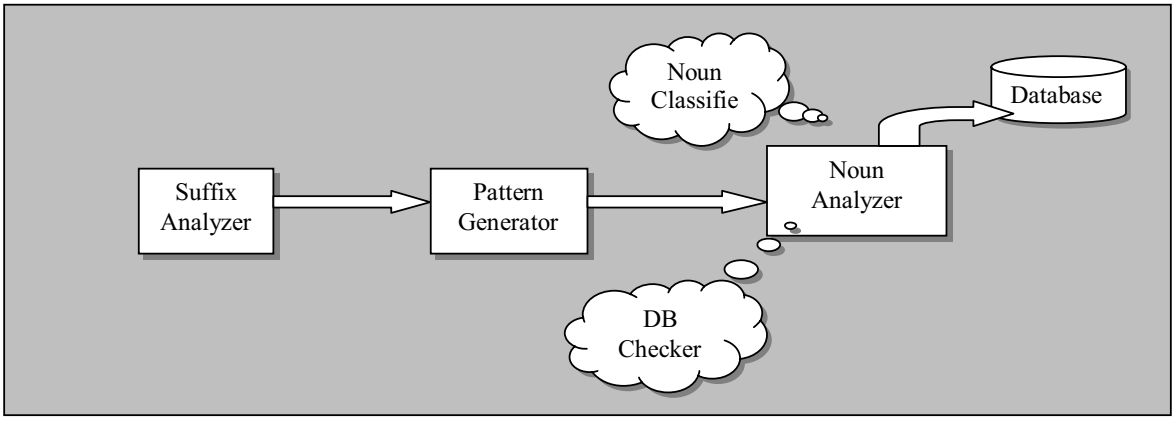

Figure 1. Noun Morphology System

4.1 Database

The database includes a Classified Noun Table that contains each root noun, source noun (singular): masculine or feminine format of the noun and the number of the group to which the

noun belongs. Each time the system identifies a new noun it adds its root to the Classified Noun Table. In addition to this table we use two tables: the Patterns Table and the Groups Table.

4.2 Noun Analyzer Module

This is the core of the system, it calls different modules and performs different tasks to identify the noun and find its paradigm. First, it passes the noun to the suffix analyzer module to drop the suffix. Second, it passes it to the pattern generator module to find the pattern. Third, it passes the noun and its pattern to the DB Checker Module to analyze the patterns and the features to see if it is a classified noun or not. If yes the module returns the name of the group to which it belongs. If the DB Checker Module cannot identify the group then the Noun Analyzer Module passes the noun to the Noun Classifier Module to classify it and identify the group it belongs to.

4.3 Suffix Analyzer Module

This module identifies the suffix, analyzes it and produces some lexical information about the noun like the number and the gender. First, it checks if any pronoun is concatenated with the noun. Second, it checks for a suffix indicating the number. Third, it checks for a suffix indicating the gender.

When the letter (ي) comes at the end of the noun there are two main cases: it could be a part of the noun so we should not drop it, or it could be an extra letter as in relative adjective, the dual noun or when the pronoun is connected to the noun and it should be dropped in this case. When the noun ends with the letters (ين), most of the time it this signifies a dual noun but some times it signifies both the plural and dual noun forms as in the following patterns: mfa9l, fa9l, mf9ull. Sometimes we have to check the pattern also to help in analyzing the suffix. Most of the time when the noun ends with (ات) this indicates it is feminine plural. To change it to singular we drop the letters (ات) and we replace them with (ة) but in few cases we just drop the letters (ات) as in (مدرجات). Most of the time when the noun ends with the letters (ان), this indicates that the noun is dual (ملعبان two playgrounds) but some times it indicates plural form as in (غزلان deer). We will handle these problems as special cases.

4.4 Pattern Generator Module

We have collected 83 different patterns including both masculine and feminine, singular and plural forms, after the suffix has been dropped see Appendix A. We used these patterns to generate a set of rules to build a finite-state machine that can used to find the pattern for any noun. The input to this module is a noun after its suffix has been dropped in the previous step, the output is one or more patterns. If more than one pattern is found we validate the string by checking the root table. The letter (م), the letter (ت) and the letter (ا) at the beginning of the noun may sometimes be the first characters of the noun root (noun: ملوك "kings" – pattern: F9wL – root: ملك), the letter (م) changed to (ف) in the pattern. Sometimes they are just extra letters (noun: مفتاح "key" – pattern: mF9al – root: فتح), the letter (م) does not change to (ف) in the pattern. We collected the nouns that begin with the letters (م), (ت) and (ا) and saved them in a file to help us to distinguish between these two cases.

4.5 Database Checker Module

This module identifies any already classified noun or any noun derived from it. It gets the noun and its pattern from the Noun Analyzer Module, finds all groups that contain the pattern, finds the singular noun (masculine or feminine) in each group and uses it to check the Classified Noun Table. If the noun appears in our tables, it fetches the number of the group to which it belongs and passes it to the Noun Analyzer to generate the results. For example the noun (مطاعم "mta9m" restaurants) has the pattern (mfa9l). This pattern appears in three different groups. See table 2. The nouns formed from these patterns have the following features. See table 3.

Table 2. The Groups of the Noun "مطاعم"

Group#	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
1	X	MF9L@	X	**mFa9L**
2	mF9L	X	X	**mFa9L**
3	**mFa9L**	mFa9L@	mFa9Lwn/ mFa9Len	mFa9Lat

Table 3. The noun of each pattern in each group

Group#	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
1	X	مطعمة	X	**مطاعم "mta9m"**
2	مطعم	X	X	**مطاعم "mta9m"**
3	**مطاعم "mta9m"**	مطاعمة "mta9m@"	مطاعمون "mta9mwn" مطاعمين "mta9men"	مطاعمات "mta9mat"

If the noun itself or any other noun derived from it has been previously classified we will find its noun root (singular noun) in the Classified Noun Table. The module will find the root (singular masculine) "مطعم" in the table and will get its group number "2" and pass it to Noun Analyzer to find the noun features.

4.6 Noun Classifier Module

This module gets the noun and its pattern from the Noun Analyzer Module. It finds all alternatives (groups) that contain the pattern and provides the user with these groups to choose one of them to reduce the number of alternatives to one. Example:
Input: The noun (مشاريع "mashare9" projects) The pattern: مفاعيل mFa9eL

Step #1: Identify the groups: each group in this table contains the pattern "مفاعيل"

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	mF9eL	X	X	**mFa9eL**
2	mF9aL	X	X	**mFa9eL**
3	mF9wL	X	X	**mFa9eL**

Step #2: Find the noun of each pattern in each group

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	مشريع "mshre9"	X	X	**مشاريع "mshare9"**
2	مشراع "mshra9"	X	X	**مشاريع "mshare9"**
3	مشروع "mshrw9"	X	X	**مشاريع "mshare9"**

Step #3: Ask the user to pick the right one. Pick the valid singular masculine of feminine noun from the list. in this example The list has three nouns: مشروع مشراع مشريع. The user helps the system to identify the valid noun, which is مشروع in this example.

Step #4: Update the database. Add a new entry to the Nouns Table. The singular masculine noun مشروع "project" is stored in the noun field and the group's number "3" is stored the group field.

5 Results

To test our system we used nouns obtained from the newspaper *Al-Raya*, published in Qatar. We have implemented the system by using Visual Basic and MS-Access. We have tested the system on 5000 nouns. We classified 592 different nouns as either singular-masculine or singular-feminine. The Suffix Analyzer Module failed to identify 245 nouns due to the different problems maintained in Section 4.3, and the Pattern Generator Module failed to recognize 180 noun due to either missing patterns or the wrong form of patterns for foreign nouns as described in Section 4.4. These missing patterns have now been added. The suffix analysis problem is hard to correct because it arises from underlying ambiguities. If the noun has been classified previously the system does not have any problem in identifying it and identifying any noun derived from it. The Noun Classifier Module found most of the nouns that the Database Checker Module failed to identify. The system identified 4640 noun 92.8% of them, 88.6% by using the Database Checker Module and the Classified Noun Table and 11.4% by using Noun Classifier Module. The system failed on 7.2% of the tested.

6 Conclusion

We have built a learning system that utilizes user feedback to identify the nouns in the Arabic language, obtain their features and generate their features with respect to number and gender. We tested the system on 5000 nouns from newspaper text. The system identified (4640) 92.8% of them, 88.6% by using the Database Checker Module and the Classified Noun Table and 11.4% by using Noun Classifier Module. The system failed on 7.2% of the tested nouns. The system classified and added 528 new different nouns to the database.

References

Al-Shalabi, R. and Evens, M., 1998. "A Computational Morphology System for Arabic". Workshop on Semitic Language Processing. COLING-ACL'98, University of Montreal, Montreal, PQ, Canada, Aug 16 1998. pp. 66-72.

Beesley, K. and Karttunen, L., 2000. "Finite-State Non-Concatenative Morphotactics". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, Oct 1-8, 2000. pp.191-198.

Roeck, A. de and Al-Fares, W., 2000. "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, Oct 1-8, 2000. pp.199-206.