Morphological Analyzer of Arabic Words Using the Surface Pattern

Said Iazzi¹, Abdellah Yousfi², Mostafa Bellafkih¹, Driss Aboutajdine¹

¹ Laboratory GSCM-LRIT, FS, Agdal-Rabat, Maroc
² Team ERADIASS, FSJES, Souissi-Rabat, Maroc.

Abstract
The present article introduces a system for the morphological analysis of the Arabic language. The system is based on the surface patterns of Arabic words. Our work in this article purports to deal with Arabic derived nouns. It is based mainly on the building of a database for the surface patterns of the latter. In order to deal with Arabic derived nouns, the article is also based on a previous study by [19] for the analysis of Arabic verbs. Our approach was tested against a corpus of 2400 Arabic words (400 verbs and 2000 derived nouns), the obtained results are very interesting and show the utility and importance of this approach.

Keywords: ATNL, Arabic Derived Nouns, Surface Pattern of Words, Morphological Analysis, Degree of Similarity.

1. Introduction

The Arabic morphological analysis is one of the tools that permit to solve the majority of the problems of Arabic. It has been extensively used in several domains of the automatic treatment of natural languages (ATNL), like documentary research, electronic dictionaries, systems of marking, etc. Several works have been realized in order to elaborate morphological analyzers of Arabic. They can be aggregated in three approaches [6] [19]:

• The symbolic approach: This approach is based on the segmentation of the word into prefixes, infixes and suffixes in order to extract the root of the Arabic word. Several morphological analyzers have been elaborated which rely on this approach [6]; [5] ; Hegazi et ElSharkawi, 1986 ; Koskenmiemi, 1983 ; Beesly, 1998 ; El-Sadany et Hashish, 1989 ; Khoja et Garside, 1999 ; Soudi, 2002). Among the well-known analyzers for this approach is that of Buckwalter. The latter consists of determining all possible segmentations of the word, looking for the results in the lists of the radicals, the suffixes and prefixes, and verifying, then, if the morphologies of each of the elements are compatible with each other while examining three tables of correspondences: prefix-radical, prefix-suffix, radical-suffix.

• The statistical approach: This approach calculates the possibilities and the probabilities that a prefix and a suffix or a radical can appear together in a data base of words (Goldsmith et John, 2001).

• The hybrid approach: this approach combines the two previous approaches (Darwish, 2002).

Among the disadvantages of these approaches, we can mention for example:

i. The dictionary of the words is very big, and it is very difficult to construct a dictionary containing all Arabic words. These dictionaries (of words) contain a sort of repetition of the nouns having the same morphological rules ("سـمـجـنـيـنـِـّاـ جـاءـَـّاـ "، "سـمـجـنـيـنـِـّاـ جـاءـَـّاـ ").

ii. These approaches use several rules at the time of morphological analysis.

To remedy these problems, we developed an independent morphological analyzer of the words dictionary, without using the rules at the time of morphological analysis. Our system uses only the surface patterns of the word to analyze.
2. Construction of the surface patterns base of the derived nouns

2.1 Derived Arabic nouns

The derived nouns are the nouns that can be derived from a verbal root. The number and the nature of these forms vary according to the status of the verb to which they are connected. Among the derived nouns (see Table 1), we can mention (Mesfar, 2008):

- **The Active Participle** "اسم الفاعل " : is a noun associated to any verb of action (transitive or intransitive) and which designates the agent of the verb, i.e. the one that has done the action. For example, the verb 'ضرب ' has the active participle ' ضرب '.

- **The Passive Participle** "اسم المفعول " : is a noun associated to any transitive action verb. It designates the patient that undergoes the action or the result of this action. For example the verb 'ضرب ' (to hit) has for passive participle ' ضرب ' (hit).

- **Verbal Noun** "المصدر " : is an abstract noun formed on the same root as the verb to which it is associated and expresses the same semantic content as the verb. A verb can have more than one verbal noun.

  For example, the verb "ود "(to like) admits four different verbal nouns "ود "، "ود "، "ود "، and "ود ".

- **The similar quality** "الصفة المشابهة " : the nouns of the similar quality indicate the absolute presence of the quality of the one who did the action, like "مغفور " 'gracious.'

- **The comparative** "اسم التفضيل " : it indicates the common quality of two nouns of which one expresses a superior degree, like "أخوف " 'more fearful.'

- **The nouns of places and times** "اسم الزمان والمكان " : they indicate the place of the time of the action, like "ملعب " 'playground.'

- **The noun of instrument** "اسم الآلة " : it indicates the means by which the action has been achieved, like "اْس " 'a spoon.'

In this article, in addition to these nouns, we have also treated the following derived nouns: "المصدر "، "اسم الهمة "， "masdar sinaai "، "masdar "، "masdar sinaai "، "masdar "， "masdar sinaai "， "masdar "، "masdar sinaai "، "masdar "، "masdar sinaai "， "masdar "، "masdar sinaai "، "masdar "، "masdar sinaai "، "masdar "، "masdar sinaai "، "masdar "، "masdar sinaai "， "masdar "، "masdar sinaai "， "masdar "، "masdar sinaai "．

<table>
<thead>
<tr>
<th>Nature and number of pronoun</th>
<th>Noun Type</th>
<th>Root Derivation</th>
<th>Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>اسم الأفعان</td>
<td>متّاً-ملّاً</td>
<td>ضرب</td>
<td>ضرب</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملزم-ملم</td>
<td>فعال</td>
<td>فعال</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملزم-ملم</td>
<td>ملازم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملازم-ملزم</td>
<td>ملازم</td>
<td>ملازم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملازم-ملزم</td>
<td>ملازم</td>
<td>ملازم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملزم-ملم</td>
<td>ملازم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم المفعول</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
<tr>
<td>اسم الفاعل</td>
<td>ملازم-ملزم</td>
<td>ملزم</td>
<td>ملزم</td>
</tr>
</tbody>
</table>

2.2 Surface pattern

The word pattern permits to detect the letters which constitute its root. The pattern of "لِمْكُومَ " is " لِمْكُومَ " . The letters "ف ع " replace the letters of the root of "لِمْكُومَ " , and the pattern of "صَارِع " is " صَارِع " (Youssef, 1999 ; Bahar, 1989–930; Hanafi, 1914 ; Zanjani, 1343).

This type of pattern cannot present the morphological variations of the word (for example the noun "نِمْل " of the verb "نِم "). That is why we proposed an adapted pattern named surface pattern [19].

The method of construction of this new pattern is the following:

If we suppose that the word whose pattern we look for is: $w = l_1 l_2 ... l_n$ ($l_i$ Character of the word $w$) and $R$ is its root.

The surface pattern of $w$ is $p = f_1 f_2 ... f_n$ with:

$$f_i = \begin{cases} f_i & \text{if } l_i \in R \\ l_i & \text{if } l_i \text{ is not in } R \end{cases}$$
And the surface pattern of the root $R = g_1 g_2 ... g_k$ ($g_i$ is a character) is $P = f'_1 f'_2 ... f'_r$ with:

$$
\begin{align*}
  f'_1 &= \text{one of the three letters "\f, 
  \h, \r" if } g_i \text{ is a consonant letter at the time of the conjugation of } R.
  \\
  f'_i &= g_i \text{ otherwise.}
\end{align*}
$$

**Example:**

The conjugation of the word "تَعَلَى" to the active participle in the 1st person singular is "تَعَلَ"; therefore, the surface pattern of the root "تَعَلَى" is "تَعَلَ" and "تَعَلَ" is the surface pattern of "تَعَلَ".

The surface pattern of "تَعَلَ" is "تَعَلَ" and of "تَعَلَ" is "تَعَلَ".

For the construction of the base of the surface patterns of the Arabic derived nouns, we treated 127 roots that represent almost all possible classes to generate Arabic derived nouns (Youssef, 1999).

Some linguists generated all Arabic derived nouns from these 127 roots, then they conjugated them to the different persons (masculine singular, masculine dual, masculine plural, feminine singular, feminine dual, feminine plural), and from these nouns, they produced the surface pattern of every derived noun.

At the end we obtained more than 6216 surface patterns which represent almost all Arabic derived nouns. (See Table 2).

<table>
<thead>
<tr>
<th>Root</th>
<th>Nature and number of pronoun</th>
<th>Noun Type</th>
<th>Surface Pattern of derived Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>فاء</td>
<td>çıkarان اسماً مفعولاً</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>فينيتاً</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
<tr>
<td>فاء</td>
<td>مفرد مذكر</td>
<td>اسماً مفعولاً</td>
<td></td>
</tr>
</tbody>
</table>

3. The approach used in our morphological analyzer

In the approach already used by [19], we noticed that for the construction of the base of the surface patterns of the verbs, it adds a phase of segmentation of the word into suffix and prefix before finding the surface pattern of this word.

Example: the word "فَوَاقِيْهِ" after the extraction of the prefix "فَ" and of the suffix "هِ" we find "فَوَاقِيْهِ". Thus, the surface pattern is "فَوَاقِيْهِ".

We look for the patterns of the word in the set of surface patterns having the same length. In this work, we were able to formulate the function that measures the similarity between the word to analyze and the surface patterns. This function has been formulated as follows:

$$
f(m, w) = \sum_{i=1}^{N} l_{m, w_i} \quad \text{with:}
$$

$$
l_{m, w_i} = \begin{cases} 
1 & \text{if } m_i = w_i, \quad (m = \hat{w}, \quad = \hat{x}, \quad = \hat{j}) \\
0 & \text{Else, and we leave the algorithm}
\end{cases}
$$

$m_i$: $i^{th}$ Character of the pattern $m$

$w_i$: $i^{th}$ Character of the pattern $w$

The function $f$ produces a set of solutions of surface patterns that we mark by $S$:

$$
S = \{ m \in P_{L(w)} \mid f(m, w) > 0 \}
$$

$P_{L(w)}$: the set of all surface patterns of $L(w)$ length.

$L(w)$: the length of the word $w$.

Example: $f(\text{فَوَاقِيْهِ}, \text{فَوَاقِيْهِ}) = 6$

$$
\begin{align*}
  f(\text{فَوَاقِيْهِ}) &= 6 \\
  f(\text{فَوَاقِيْهِ}) &= 6 \\
  f(\text{فَوَاقِيْهِ}) &= 0
\end{align*}
$$

Subsequently, for every surface pattern $m_b$ of the word $w$ we look for its roots $R_{m_b}$. To find the roots of the word $w$, we seek the positions of characters "
, 
, 
, 
" in the surface patterns of the word $w$, and extract the characters associated with these positions. The roots of the
word w are found by replacing these characters in the surface patterns of root. For example, for the word قانون فعل - فعلن we find the surface patterns:

- قانون فعل - فعلن with the surface pattern فعل for its root
- قانون فعل - فعلن with the surface pattern فعل for its root.

After the application of our method we find the two following solutions:

- قانون - فعلن - فعل
- قانون - فعل - فعلن

As the root فعل doesn't exist in Arabic, we keep only the second solution فعل ل (see Tab 3)

Table 3: Example of the results of the morphological analyzer of words

<table>
<thead>
<tr>
<th>Nature and number of pronoun</th>
<th>Type nom</th>
<th>Words pattern (w)</th>
<th>Roots pattern</th>
<th>Words Roots</th>
<th>Words (w)</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>إنت</td>
<td>متى-منزة</td>
<td>اسم-فعل</td>
<td>كافف-في-نن</td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
<tr>
<td>ملكي-ميزة</td>
<td>اسم فعل</td>
<td>كافف-في-نن</td>
<td></td>
<td>قانون - فعال</td>
<td>قانون - فعال</td>
<td></td>
</tr>
</tbody>
</table>

4. The implementation

To test our approach, we first constructed all surface patterns of the derived Arabic nouns. This stage has been achieved by linguists, and they used a set of Arabic references (Mustapha, 1999; Baharak; Hanafi et al., 1914; Zanjani, 1343).

For the implementation of our approach, we developed a program in Java which consists of the following parts (see diagram 1).

- Part 1 segment the word into suffixes, the prefixes and root.
- Part 2 look for the surface patters of the solutions given by part 1.
- Part 3 look for the roots from all surface patterns returned by part 2.

Part 4 verify the validity of these roots while verifying if they exist in the base of the roots or not. This approach has been tested on 2400 words (400 verbs and 2000 derived nouns). These words are different from those used in the phase of surface patterns construction.

The global error rate found is 3.9%. The majority of these errors spring mainly from the insufficiency of the surface patterns data base. There are some derived nouns whose surface patterns don't exist in our patterns base. For the rest of the errors, they come from the phase of generation or the construction of these surface patterns.

5. Conclusion:

Our contribution was the treatment of Arabic derived nouns which have not been treated in the case of [19]. Afterwards, we formulated the function that measures the similarity between the words and the surface patterns. Moreover, we could reduce the size of the patterns base while eliminating the phase of prefixes and suffixes addition to the surface patterns. The error rate found is reasonable, and shows the interest of our approach. Subsequently, we will treat the non-derived nouns.

References


