ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

The morphological analysis of Arabic verbs by using the surface patterns

A. Yousfi
The institute of studies and research for the arabization
University Mohamed V Souissi, Rabat, Morocco

Abstract

In this article, we present a new system of morphological analysis using the surface patterns of the word to be analyzed. This approach required in the initial phase a classification of all Arabic verb roots by using the morphological rules of these last, and the second phase is the construction of data base of conjugated surface patterns.

This system is capable to analyze all Arabic verbs by decomposing the word to the prefixes, suffixes and roots. This approach has been tested on a corpus of 4000 verbs, the results were encouraging, and the error rate is 4%.

Keywords: *Arabic verbs, Morphological Analysis, Surface pattern, degree of similarity.*

1. Introduction

The morphological analysis has been widely used in several fields of Natural Language Processing (NLP) such as Information Retrieval (IR) systems, data mining, electronic dictionary, tagging systems and language model.

The several works have been done in the area of the Arabic morphological analysis. The three approaches are used in these works [1]:

- The Symbolic approach: this approach is based on removing all attached prefixes and suffixes in an attempt to extract the root of a given Arabic surface word. Several morphological analyzers have been developed, e.g. [2], [3], [4], [5], [6], [7], and [8].
- The statistical approach: this approach computes the probability that a prefix, a suffix, or a template would appear from a corpus of surface word [9].
- The hybrid approach: this approach uses combination rules of morphemes with statistics. This method is based to build morphological rules and to estimate the probability that a prefix, a suffix, or a template would appear [1].

These approaches have the disadvantages, one mentions for example:

- The dictionary of stems is very large and it is very difficult to build a dictionary containing all Arabic stems. These dictionaries (of stems) contain a sort of repetition of the verbs having the same morphological rules ("جَمَعَ - يَجْمَعُ"، "فَتَحَ - يَفْتَحُ").
- These approaches use several rules at the time of the morphological analysis.

To remedy these problems, we have developed a morphological analyzer independent of stems dictionary and don't use the rules at time of the morphological analysis. Our system uses only the surface patterns of the word to be analyzed.

2. Classification of verbs by using the surface patterns

The word pattern permits to detect the letters of the root of the word. The pattern of "يَكْتُبُونَ" is "يَفْعَلُونَ", the letters " ف،ع،ل " replace the letters of root of "يَكْتُبُونَ", and the pattern of "نَالَ" is "فَعَلَ" [10].
This type of pattern cannot present the morphological variations of the word, it is why we proposed an adapted pattern called surface pattern.
The construction method of this new pattern is:
If one supposes that the word to find his pattern is:
$w = l_1 l_2 ... l_n$ (l_i is letter), and R is her root.

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

$$\begin{cases} f_i \text{ is one of the three letters " ف،ع،ل " if } l_i \text{ is in R} \\ f_i = l_i \text{ if } l_i \text{ is not in R} \end{cases}$$

And the surface pattern of the root $R = g_1 g_2 ... g_k$ (g_i is letter) is $p = f'_1 f'_2 ... f'_k$, with:

$$\begin{cases} f_i' \text{ is one of the three letters " ف،ع،ل " if } g_i \text{ is a steady letter at the time of the conjugation of the R.} \\ f_i' = g_i \text{ else} \end{cases}$$

Example:

The conjugation of verb "أَخَذَ" to the imperfect and 1st person, is "آخُذُ", then the surface pattern of root "أَخَذَ" is "أعَلَ" and "آعُلُ" of "آخُذُ".

The surface pattern of "نِلْتُ" is "فِلْتُ" and "فَالَ" of "نَالَ"

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

For the construction of data base of surface patterns of Arabic verbs, we classified more than 9800 root according to the conception of the surface pattern. We got more 100 surface patterns (Table1) of these roots [11].

Classe	Surface pattern
7	أعَلَ
8	أعْلَ
9	أعِلَ
20	آعَى
70	اِسْتَفْعَأَ
66	اِنْفَعَلَ
10	فَالَ
1	فَعَلَ
2	فَعُلَ
3	فَعِلَ
50	فَعْلَلَ

Table1: a part of the surface patterns of roots

3. Construction of data base of surface patterns of verbs

For the construction of data base (DB) of all surface patterns of Arabic verbs, we developed a program writes in JAVA to conjugate the 100 surface patterns of roots to all tenses (8 tenses) and all persons (13 persons and 5 persons of imperative tense (الأمر)).
The number of the conjugated surface patterns in the data base is: 100*8*13+100*5=10900.

In order to eliminate the decomposition phase of the word to the prefixes, suffixes and roots (during the morphological analysis), we attached all prefixes and suffixes to the conjugated surface patterns of DB (the total number of surface patterns in DB is the order of 350000).

The DB contains other information of types: tense, person, surface pattern of root … (Table2).

Surface pattern of root	Person	Tense	Surface pattern
فَعَلَ	هو	المضارع المعلوم المرفوع	سَيَفْعَلُهُ
فَعُلَ	هو	المضارع المعلوم المرفوع	سَيَفْعُلُهُ
فَالَ	أنا	الماضي المعلوم	وَقُلْتُ
أعَلَ	أنا	المضارع المعلوم المرفوع	آعُلُكَ
…	…	…	…

Table2: an example of a part of the DB.

4. The approach used in our analyzer morphological

Let *w* the word to be analysed and $E = \{p_1, p_2, ..., p_N\}$ is the DB of the surface patterns.
Our approach consists to finding all surface patterns of *w*, by using the degree of similarity *s* between *w* and surface pattern, *s* permits to count the number of letters that repeated in *w* and in surface pattern. The set of possible surface patterns of *w* is:

$$Sol = \{ p \in E \mid s(p, w) > 0 \}$$

We indexed the DB by using *s*; the research of surface pattern gets only in a query of size inferior to 20. This indexation decreases the duration of research in the DB. After the obtaining the set *Sol*, we get the information associated to these surface patterns (tense, person, surface pattern of root …).

Example:

The morphological analysis of the word "فَأعْلَمَهُنَّ" is achieved by the following stages:

- The system search the surface patterns of "فَأعْلَمَهُنَّ", it finds "فَأفْعَلَهُنَّ". This surface pattern is repeated three times in our DB :
  - * فَأفْعَلَهُنَّ ماضي-معلوم هو أفْعَلَ
  - * فَأفْعَلَهُنَّ مضارع-معلوم-منصوب أنا فَعِلَ
  - * فَأفْعَلَهُنَّ مضارع-معلوم-منصوب أنا فَعَلَ
- The system uses these solutions to extract the roots associated to these solutions. Its finds the following roots: "أعْلَمَ"، "عَلِمَ"، "عَلَمَ".
- The system verifies if these roots exist in the base of the Arabic roots.
- To the final phase the system presents the just morphological analysis (Fig. 1).

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

Fig. 1: The interface of our morphological analysis program.

Among the advantages of our system, we mention:
- The phase of decomposition of word (at the time of the morphological analysis) in prefix, a suffix and root is eliminated. This operation will decrease the duration of the morphological analysis.
- Our system doesn't use the morphological rules at time of the morphological analysis.
- Our system doesn't use the dictionaries of stems.

5. Implementation

The evaluation of our approach is made by the realization of a program in java. This program is constituted of four packages:

- Package of conjugation the surface patterns of roots to all tenses and all persons.
- Package that permits to attach the different suffixes and prefixes to the conjugate surface patterns.
- Package of indexation of data base of all surface patterns.
- Package of morphological analysis. This package is constituted of two modules (Fig. 2), Module1 permits to find the different surface patterns of word (Input), Module2 extracts all possible solutions and verify the validity of these last.

This approach has been evaluated on 4000 verbs, the error rate is 4% and the duration of these verbs is 10 seconds (2,5ms for each verb).
The majority of these errors (4%) are mainly that the entered words for the analysis are not the associated surface patterns.

6. Conclusions

In this article, we presented a new approach of morphological analysis. This approach has been applied on the verbs.

This approach remains always true for the derivative names [12], [13] it is sufficient to construct date base of surface patterns of derivative names.

This data base of surface patterns of derivative names is under construction by the help of the linguists.

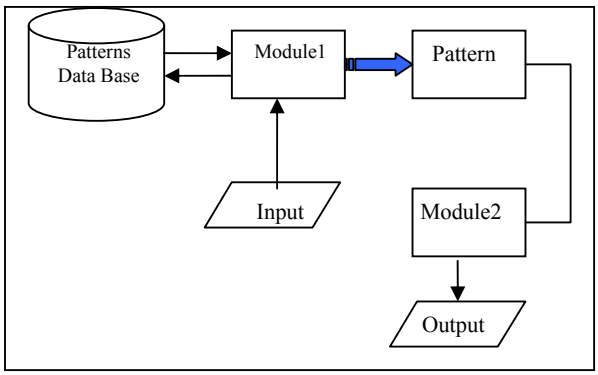

Fig. 2

7. References

[1] Darwish, K. : "Building a shallow Arabic morphologi-cal analyser in one day" in Proceedings of the ACLWorkshop on Computational Approaches to Semitic Lan-guages, Philadelphia, PA, 2002.

[2] Buckwalter, T. : 2002, Buckwalter Arabic Morphological Analyzer. Version 1.0. Linguistic Data Consrtium, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.

[3] Hegazi, N., and ElSharkawi, A. A. 1986. Natural Arabic Language Processing, Proceedings of the 9th National Computer Conference and Exhibition, Riyadh, Saudi Arabia, 1-17.

[4] Koskenniemi, Kimmo. Two Level Morpology: A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dep. of General Linguistics, University of Helsinki, Helsinki, 1983.

[5] Beesly KR: Arabic Morphology Using Only Finate-State Operations, Proceedings of the Workshop on Computational Approaches to Semetic languages. Montreal, Quebec, pp 50-57.

[6] El-Sadany, T. A., and Hashish, M. A. 1989. An Arabic Morphological System. IBM Systems Journal. Vol.28, No.4, 600-612.

[7] Soudi, A.: 2002, A Computational Lexeme-Based, Treatment of Arabic Morphology. Doctorat d'état, Mohamed V University.

[8] Khoja S., Garside R.. Stemming Arabic text. Computer Science Departement, Lancaster University, Lancaster, UK.

ISSN (Online): 1694-0784
ISSN (Print): 1694-0814

[9] Goldsmith, J.A. (2001). Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, 27:2 pp. 153-198.
[10] Mustapha G., الشيخ مصطفى الغلاييني، جامع الدروس العربية. المكتبة العصرية، 1999
[11] Sam A, Youssef D., سام عمار، يوسف ديشي، مجموعة بيشرال الأفعال العربية. هاتييه باريس، 5 أكتوبر 1999
[12] Yousfi, A., Sabri, S., Bouyakhf, E., Système d'analyse morphologique des noms Arabes. JETALA (Journées d'Etudes sur le Traitement Automatique de la Langue Arabe), 2006, Rabat 5-7 juin, 2006.
[13] Yousfi, A., Sabri, S., Bouyakhf, E., Système d'analyse morphologique des noms Arabes. MCSEAI'06 December 07-09, 2006, Agadir.