

A Morphological Analyser for Arabic Language.

Mourad MARS

LIDILEM laboratory
University of Stendhal,
Grenoble3, FRANCE.
Mourad.Mars@e.u-grenoble3.fr

Georges ANTONIADIS

LIDILEM laboratory
University of Stendhal,
Grenoble3, FRANCE.
Georges.Antoniadis@u-grenoble3.fr

Mounir ZRIGUI

UTIC laboratory
Faculty of Sciences of Monastir,
TUNISIE.
Mounir.zrigui@fsm.rnu.tn

Abstract

Morphological analysis is an essential component in language engineering applications. For the Arabian language, the omnipresent agglutination phenomenon makes delicate the creation of the morphological analyser and needs more treatments and specific tools. In this paper, we present the creation of a morphological analyser for Arabic language. It will be integrated in the MIRTO platform developed in LIDILEM laboratory to create educational activities for the training of Arabic.

For realization of our morphological analyser we achieved three labelled dictionaries, a basis of rules for reconstitution and restoration of the good stems, a basis of rules to assign a final label to the graphic word as a whole and a robust word segmentation algorithm for Arabic language. Moreover, at the time of the carving we use a basis of morphological classes, developed at LIDILEM, to get a final validation of our decomposition.

1. Introduction

The morphological analysis of Arabic is interested, as the other languages, to the structure of the word. But being given the wealth of the Arabic word's structure and the problem of agglutination, the operation becomes more difficult than for the other languages. In what follows, we describe an original algorithm of morphological analysis achieved while exposing the different stages of analysis and we present all used resources.

2. Works in the domain

Many works have been led during these last years and new theories were born. Therefore, many systems have been achieved in the domain of the morphological analysis.

- The works of Youssef Tahir, Nouredine Chenfours and Mostafa Hartis [2004]:

Their work moves toward the conception and the organization of a linguistic data base for the Arabian language. This data base includes most primitive linguistics of Arabic [Youssef T, Nouredine C, Mostafa HS, 2004].

- Ken Beesley: The Xerox Arabic Morphological Analysis and Generation ":

This morphological analyser use the XEROX" finite developed state technology". It treats the unvoyeled, voyeled and semi-voyeled words. A trial version is on the site of XEROX www.xrce.xerox.com. The major inconvenience, this morphological analyser is not free.

- The Morphological sensor of Sakhr software:

The Multi-Mode Morphological Processor (MMMP) Sakhr is a morphological sensor; it provides the analysis of basis for all Arabian word. This sensor covers the whole modern and classic Arabic language. Unfortunately, there is not a trial version to check it.

3. Resources for a morphological analyser for Arabic language

One of the main stages of the realization of a morphological analyser for the Arabian language is the conception and the organization of a linguistics database. The database achieved includes the majority of the primitive linguistics of Arabic as the verbs, the names, the particles, the pre-stems and the post-stems.

3.1. Dictionary of Arabic lexicon

In this dictionary, we have three categories of words: the verbs, the names and the particles,

3.2. Dictionary of Pre-Stems

The pre-stem are the result of the concatenation of the proclitics with the compatible prefixes. To achieve this dictionary, we use a set of the compatibility rules.

After the application of the rules of concatenation that we achieved, the total number of the labeled Pre-Stems is 88.

Example:

Label	Pre-Stem
أ الاستفهام+ف العطف+ت مض	أفت

3.3. Dictionary of Post-Stems

This dictionary contains 233 labeled post-stems. They are gotten by the combination between suffix and enclitic(s).

Label	Post-Stem
ماض+ض+3مذ+جم+ضم+ض+3مذ+جم	تهم

4. Word Morphology Analysis algorithm

4.1. Extraction of the pre-stems

We apply this procedure to the graphic word: وبالأمل /wabelamal/ "and with the hope", we find this list "و/wa/", "وب/ wabe/" and "وبال/ wabel/". We suppose usually l'existence of the pre-stem "".

4.2. Extraction of the post-stems

The process is precisely the same that for the pre-stems, the minor difference that this time we start the extraction from the end of the word to list all post-stems of the word.

4.3. Extraction of the stems (the use of morphological rules)

The execution of the modules of extraction of the post-stems and pre-stems gives us at the exit the sets of the post-stems and possible pre-stems that are in the word.

For each couple of pre-stems and post-stems, we extract the stem and we verify if it is a valid stem in Arabic.

The deletion of the pre-stem and the post-stem is not sufficient, the concatenation of some pre-stem and post-stem generate some transformations that must remain reversible, that pushed us to achieve a basis of morphological rules of transformations for Arabic language.

Example of rule: If the pre-stem contains ل /le/ and the first letter of the stem is the « َ » gemination. In this case, the deletion is not sufficient, it is necessary to replace it by the letter ل/ل.

Example: للعب /lella''eb/ ⇒ ل + ع + ب ⇒ لل+لعب (After application of the rule).

4.4. Rules and labeling

After the extraction of the different part, we verify the compatibility of every triplet pre-stems, stems and post-stems by checking in the basis of rules. This basis is structured on morphologic class, each class regroup all elements having the same morphologic property. There are classes for pre-bases and classes for post-bases.

The application of these rules generates the elimination of all invalid decomposition.

6. References

- [1]- Joseph Dichy, "Morphosyntactic specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects", Proceedings of the ACIDA' 2000 conference, Monastir (Tunisia), 22-24 March 2000, Corpora and Natural Language Processing vol., 55-60.
- [2]- Beesley K, "Arabic Finite-State Morphological Analysis and Generation". Proceedings of COLING-1996, pages 89– 94.
- [3]- Darwish, K. "Building a Shallow Arabic Morphological Analyzer in One Day", Proceedings of the Workshop on Computational Approaches to Semitic Languages, 2002.
- [4]- Mourad MARS, Mohamed BELGACEM, "Developed of a morphological analyser for arabic language, tool for creation of educational activities of training of Arabic", Workshop "TEL in working context", 13-15 November 06, Grenoble, France. 2006.
- [5]- Riadh Ouersighni, "A major offshoot of the DIINAR-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts", ENSSIB, 2000.
- [6]- Youssef Tahir, Nouredine Chenfour et Mostafa Harti, "Modélisation à objets d'une base de données morphologique pour la langue arabe", JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, 20 avril 2004.