A COMPRESSION TECHNIQUE FOR ARABIC DICTIONARIES : THE AFFIX ANALYSIS.

Abdelmajid BEN HAMADOU

Département of computer science -FSEG Faculty
B.P 69 - Route de l'aéroport -
SFAX - TUNISIA

ABSTRACT

In every application that concerns the automatic processing of natural language, the problem of the dictionary size is posed. In this paper, we propose a compression dictionary algorithm based on an affix analysis of the non diacritical Arabic.

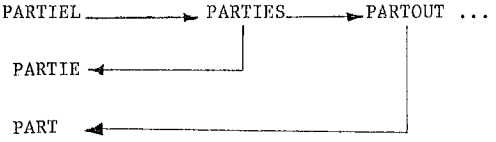
It consists in decomposing a word into its first elements taking into account the different linguistic transformations that can affect the morphological structures.

This work has been achieved as part of a study of the automatic detection and correction of spelling errors in the non diacritical Arabic texts.

I- INTRODUCTION

In every application that concerns the automatic processing of natural language, the problem of the dictionary size is posed. We can approach this important question in several ways and particularly :

- By grouping together the common prefixes of the different language words. In the PIAF system,(interactive program for French Analysis) for instance, words are represented in chained lists following an alphabetical order [COUR 77]



- By creating multiple dictionaries: one for each major topic area. This approach requires, in addition, a common base dictionary. When a particular area is concerned, a temporary master dictionary is created by increasing the base dictionary with selected local ones.

- By using the Affix analysis which consists in performing a morphological analysis in order to identify, in a given word, the redundant elements (Affixes). The dictionary will be limited to the non-redundant elements (roots). This technique is used specially in the DEC10 - SPELL system for detecting and correcting spelling errors.

In the present paper, we will develop this last approach for the non_diacritical Arabic.

The particularities of the algorithms that we propose, stem, in great part, form the specificities of the language used :

- Words are written in consonantic form
- Words can contain infixes
- Morphological structures can be altered by linguistic transformations.

This work has been developed within a national research project for the study of the automatic detection and correction of spelling errors in Arabic texts [BEN 86]

II - THEORETICAL ASPECTS

Let V be a finite Set
and V^*, the set of words built on V including nul sting noted $\emptyset$

$W \in V^* \quad W = W_1 W_2 \ldots\ldots W_n. \quad W_i \in V \quad i \in [1,n]$

let $V^+ = V^* - \{\emptyset\}$

1°/ Prefix (W)
Let $W = W_1 W_2 \ldots\ldots W_n. \; W \in V^+$
We call order i prefix the quantity $P_i = W_1 W_2 .. W_i \quad (1 \leq i < n-1)$
the order o prefix is $\emptyset$

2°/ Suffix (W)
Let $W = W_1 W_2 \ldots\ldots W_n \quad W \in V^+$
We call order j suffix the quantity $S_j = W_j W_{j+1} \ldots\ldots W_n \quad (1 < j \leq n)$
the order n + 1 suffix is $\emptyset$

3°/ Infix (W)
Let $W = W_1 W_2 \ldots\ldots W_n \quad W \in V^+$
We call order 1 infix the quantity $I = W_k \quad (1 < k < n)$
We call order 2 infix the quantity $I = W_k , W_1 \quad (1 < k < 1 < n)$
the order zero infix is $\emptyset$

4°/ Root (W)
Let $W = W_1 W_2 \ldots\ldots W_n \quad W \in V^+$
We call Root the quantity : $R = W_p \ldots\ldots W_q \quad (1 \leq p < q \leq n), \; (\text{card}(R) \leq q-p+1)$

5°/ Card ($\mathcal{P}_i$)
Let $W = W_1 .. W_n \quad W \in V^+$
Let $\mathcal{P}_i = \{\emptyset , P_1, P_2, P_3, \ldots P_i\}$
Card ($\mathcal{P}_i$) = i + 1 if i ⩾ 1
Card ($\mathcal{P}_i$) = 1 if $\mathcal{P}_i = \{\emptyset\}$

6°/ Card ($\mathcal{S}_j$)
let $W = W_1 \ldots\ldots W_n. \; W \in V^+$
let $\mathcal{S}_j = \{\emptyset , S_j , S_{j+1}, \ldots , S_n\}$
- Card ($\mathcal{S}_j$) ⩽ n-j+2 if $(1 < j < n)$
- Card ($\mathcal{S}_j$) = 1 if $\mathcal{S}_j = \{\emptyset\}$

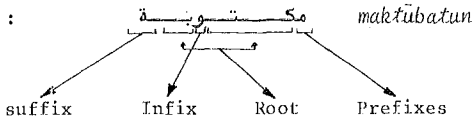
III - AFFIX ANALYSIS

1. Morphological decomposition

The Affix analysis consists in decomposing a given word into its first elements among which we can distinguish the affixes (prefix, infix and suffix) which are the redundant elements of the language and the root which is its non redundant one .

This decomposition is based on the derivational structure of the language : nearly all the words are obtained by adding an affix combination to a given root.

Ex : مكتوبة *maktūbatun*



- Root = كتب *kataba*
- Prefix = م *m*
- Infix = و *w*
- Suffix = ة *t*

Among the possible affix combinations, we distinguish those that are valid and those that are not. Valid combinations constitute what is called Morphological Pattern (M P)

For a given word, the number of possible morphological decompositions depends on the root, according to whether or not it contains characters which can be assimilated to different affixes.

This number is calculated using the following formula :

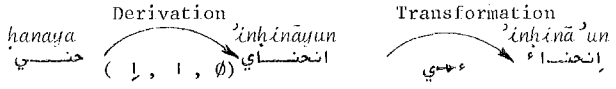
$$Nd = Card(\mathcal{P}_i) \cdot Card(\mathcal{S}_j)$$

2. Study of the morphological transformations

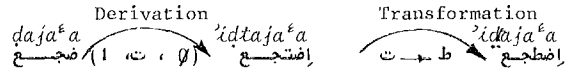
The morphological derivation for a root can be accompanied with transformations caused by linguistic phenomena such as assimilation, contraction, metathisis.

These transformations can affect the Root as well as the affixes (M P). The Roots affected are mainly those which contain the characters yaa : ي , Waw : و and hamza : أ

EX 1 : Root affected.
Consider the root : حنـــي and the MP =(إ , ا ,Ø)



EX 2 : Affix Affected.
Consider the root : ضـــجـــع *daja$^{\varepsilon}$a* and the MP =(إ , ت , Ø)

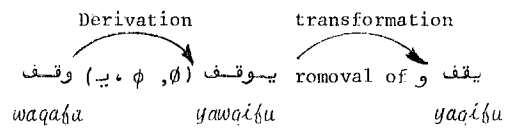


The morphological transformations can be classified into two categories :

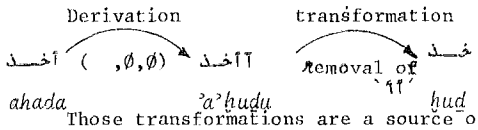
- The morpho-phonological transformation are those that substitute a character for another one without changing the length of the word (isometrical transformations)-(see EX1 and EX2).

- The purely phonological transformations are those that suppress one or more characters, therefore they modify the length of the word.

EX 3 : consider the root وقـــف



EX 4 : consider the root أخـــذ



Those transformations are a source of ambiguity for the morphological decomposition. To remove these ambiguities, we use heuristics among which we can mention for instance :

Let D be the morphological derivation operator such as : $D(R, P, I, S) = W \quad W \in V^+$

and T the operator composed of a derivation followed by a transformation. And D^{-1} the morphological decomposition operator (inverse of D) and T^{-1} the morphological decomposition operator taking into account the transformational rules (inverse of T).

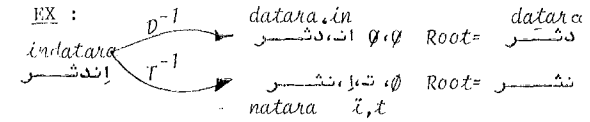
Consider W the word to be analysed.

If $D^{-1}(W) = (R_1, P_1, I_1, S_1), R_1 \in V^+$ and $P_1, I_1, S_1 \in V^*$

and $(W) = (R_2, P_2, I_2, S_2)$, $R_2 \in V^+$ and $P_2, I_2, S_2 \in V^*$

So R_1 is the selected root (R_2 is rejected)

EX :



The root retained is : دشـــر *datara*

This heuristic means that the transformations can not be done at the expense of semantics.

IV - IMPLEMENTATION :

The affix analysis is composed of two modules (See Fig. 1) :

- morphological decomposition module
- validation module

1. The morphological decomposition module permits to identify the different affix combinations. It is executed in two steps :

Step one : Identification of prefixes and suffixes by using a table of prefixes and a table of suffixes.

Step two : Identification of the infix by anaysing the remaining chain after eliminating P and S.

The analyser has a single initial state and as many ways as there are infix possibilities.

The interest of realising this decomposition into two steps lies in the use of a single analyser in order to recognise all the morphological forms. we distinguish different morphological Patterns .

2. Validation module

The two preceding steps lead to a list of candidate decompositions.It is necessary to apply to this list an adequate validation mechanism to sort out the valid decomposition

This filtering can provide multiple solutions. In these conditions, we talk about morphological ambiguity that can not be removed without considering the context of the word in the sentence.

However, the affix analysis used for the purpose of verifying whether or not a word belongs to the language can be content with the first valid decomposition.

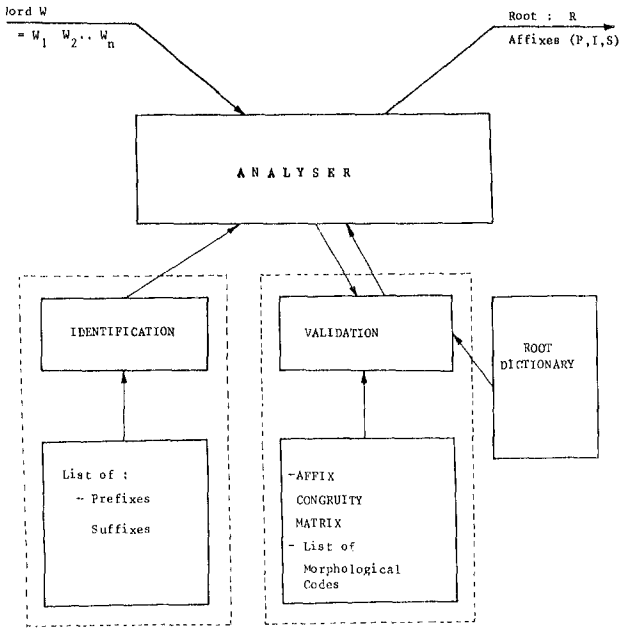


Fig 1 : Functional Affix Analysis diagram

The validation is based on the principle of affix congruity and on the result of the root dictionary checking.

The affix congruity arises at three different levels :
- Compatibility between the prefix (P) and the suffix (S)
- Compatibility between the couple (P,S) and the infix (I).
- Compatibility between the Morphological Pattern (P,I,S) and the Root (R) .

The compatibility between P and S is obtained from the affix congruity matrix C (P_i, S_j) composed of 609 elements (21 prefixes and 29 suffixes). The values attributed to a couple (P_i, S_j) are :

$C(P_i, S_j) = 0$ if P_i and S_j are incompatible

$C(P_i, S_j) = N_k$ if P_i and S_j are compatible, $N_k \in [1, 226]$

The compatibility between (P_i, S_j) and the infix I_k is obtained by performing the intersection of the Morphological Code (MC) generated by the analyser with the set of Morphological Codes associated with the couple (P_i, S_j) . This set or list is referred to by N_k. Let L be this list $L = \{MC_1, MC_2, \ldots MC_l\}$

If $MC \wedge MC_i = 0$ So (P_i, S_j) and I_k are incompatible $i \in [1, l]$

If $MC \wedge MC_i = MC$ So (P_i, S_j) and I_k are compatible $i \in [1, l]$

The Compatibility of the Morphological pattern with the root does not have a morphological origin but it is essentially of a semantic one.

EX : The Word استأكل *ʾistaʾkala* does not exist because the root أكل *ʾakala*

and MP = (است, Ø , Ø) are incompatible. *ʾist*

The detection of this incompatibility requires flagging the dictionary for each root with its legal non-systematic morphological patterns (ex : derived verb forms ,'masdar' , same nouns).

The dictionary look-up permits to verify whether the word analysed belongs to the linguistic corpus or not. It plays a decisive part in identifying the valid root if the analysis, for one morphological pattern, generates several candidate roots. (nondeterministic analysis).

EX : Consider the root and MP =(ا ، ت . Ø)

Derivation Transformation

أخذ *ʾahada* (ا، ت، Ø) إتخذ *ʾiʾtahada* r3 إتّخذ *ʾittahada*

The decomposition of the target word اتخذ according to the transformation rules gives the three plausible roots : أخذ *ʾahada* وخذ *wahada* تخذ *tahada*

These transformation rules are the following ones:

r_1 : *ʾitta* اتت ⟶ *ʾitta* اتّ

r_2 : *ʾiwta* اوت ⟶ *ʾitta* إتّ

r_3 : *ʾiʾt a* إأت ⟶ *ʾitta* إتّ

The dictionary look-up enables to suppress the candidates : تخذ *tahada* and وخذ *wahada*

Our root dictionary being used has been built by taking census of the roots related to the linguistic corpus of the Maghreb Countries. This corpus has been done by the Permanent Commission of Functional Arabic [P C F A 76]

The size of the obtained dictionary is about 1,500 three-character roots and 100 four - character roots Its increase can easily be done thanks to its evolutionary structure.

Acces to this dictionary is direct. The access argument is calculated from the first three characters of the root and its length L.

V - CONCLUSION

The affix analysis permits to replace an important dictionary containing roots only. This technique has proved efficient for Arabic because of its derivational structure. We have tested this technique on a corpus made up of 100,000 words or so using the dictionary of the 1,600 roots.

The programs are written in FORTRAN for reasons of portability, easy calculation of the Dictionary access argument and index manipulation.

Used in the context of the detection and correction of spelling errors, the affix analysis is interesting in that :

- It makes easier the use of the dictionary loaded in memory
- It performs a natural cutting of the words, which facilitates the algorithms of automatic correcting based on inferential mechanisms and heuristics

These features give the suggested algorithms some originality and a contribution to the work in the field of Arabic morphological analysis.

BIBLIOGRAPHY

(BEN 86) - A. BEN HAMADOU : Automatic detection and correction of spelling errors in Arabic texts. 2nd International Baghdad conference 24-26 March 86.

(COUR77) - J. COURTIN : Algorithmes pour la traitement interactif des langues naturelles.-Th.Etat GRENOBLE 77

(WOOD70) - W.A. WOODS : Transition Network grammar for natural language analysis C.A.C.M VoL 13 N° 10 oct 70.

(PC FA 76) - Permanent Commission of Functional Arabic L'arabe Fonctionnel. 2nd Edition - Tunis 1976.