The Use of the Relational Concept in the Arabic Morphological Analysis

Said Iazzi¹, Abderrazak Iazzi²
LRIT Associated Unit to the CNRST-URAC29, Faculty of Sciences, Mohammed V University Rabat, Morocco

Saida Laaroussi³
Laboratoire des Sciences de l'Ingénieur, ENSA, IbnTofail University, Kenitra, Morocco

Abdellah Yousfi⁴
Team ERADIASS FSJES, Mohammed V University Rabat, Morocco

Abstract—The Arabic language differs from other natural languages in its structures and compositions. In this article we have developed an Arabic morphological analyzer. For this, we have used the relational concept in the database to build our Arabic morphological analyzer. This analyzer uses a set of tables which are linked together by relationships. These relations model certain numbers of compatibility rules between different affixes. Our morphological analysis have been trained and tested on the same databases. The tests of our new approach have given good results and the numbers obtained are very close to those of existing analyzer.

Keywords—Arabic language rules; morphology; morphological analyzers; database; relational concept

I. INTRODUCTION

Morphological analysis is a central task in language processing. It consists in detecting the different morphological entities of an input word and provides a morphological representation of it. Morphological analysis has been and remains the focus of researchers in the automated processing of the Arabic language [1][6][9][10][18].

Studies on Arabic morphology at the computer level have received great attention from computer engineers and linguists since the early eighties. A large number of morphological analyzers are designed for use in various applications. The attention is due to the richness and the complexity of Arabic morphologies, the importance also appears for the morphological analyzers in the main applications to facilitate and provide solutions in the fields of machine translation, information search and information retrieval [3][4][8][21].

Automatic language processing requires several efforts in the development level of all advanced computer methods and techniques [5][7][15][23]. For many automatic language processing tasks, a complete and rich lexical database is essential, even a simple word list can often be an invaluable source of information. One of the most difficult problems with lexicons is that of non-vocabulary words, especially for languages that have a richer morphology like Arabic. To evaluate our morphological analysis systems, we need a body adapted to Arabic morphological analyzers that facilitate data processing tasks and have efficient results. Corpus analysis is focused on the experimental, while interpretation can be qualitative or quantitative.

We also describe the construction and the methodology of the necessary linguistic resources, a morphological dictionary and an adapted morphological corpus, and assess the effect of resource size on the accuracy of the analysis, showing what results can be obtained with limited linguistic resources [12].

II. APPROACHES USED IN MORPHOLOGICAL ANALYSIS

There are different ways to build morphological analyzers. There was research to set the rules of grammar, morphology, grammar and spelling to build morphological analyzer systems. The development requires the study of the properties of the data words by essentially raising issues concerning the morphological analysis and presentation of Arabic words [9][10][14][15].

The methods used in the construction of morphological analyzers are quite varied. Indeed, some researchers have developed methods based on finding diacritic symbols at the character level, others have exploited these methods to identify diacritics at the word level. A group of researchers has developed hybrid methods coupling approaches to improve these methods. Darwish [15] suggested classifying the approaches into the symbolic approach, the statistical approach and the hybrid approach.

Researchers in the field of morphological analyzer use several methods to analyze a word:

A. Approaches based on Linguistic Rules

For Arabic morphological analyzers several researchers use approaches based on linguistic rules. They use a knowledge base of rules written by linguists to assign solutions to different morphological attributes of Arabic words. The approach based on linguistic rules uses algorithms purely based on the morphological knowledge of the language. It requires rules to cover all morphological shapes. These rules are often classified into grammatical, structural and logical categories. This consists of using criteria and linguistic properties in the form of rules expressing the functioning of the natural language used.

The linguistic approach requires a large number of lists and tables. To develop a set of rules to find the appropriate decomposition, this approach is based on a thorough morphological analysis of the Arabic language [2][14].
The subjective linguistic approach simulates the process used by a linguistic expert. It consists of removing affixes by comparison with predefined lists and transforming what remains, the stem into the root, after a possible alteration by the addition, deletion or modification of some of its letters.

B. Dictionaries based Approach

The resources used for a morphological analyzer are a dictionary of root words that has been created manually using different resources. In addition, a morphological dictionary is used for both a morphological analyzer and a morphological generator, depending on the direction in which it is read by the system [14].

C. Approach based on Patterns of Words

The use of morphological patterns, depending on morphological affixes in all its forms, are patterns for verbs, patterns for names, patterns for adjectives, etc. There are some common patterns between these types.

To apply this type of analysis, it was necessary to determine the morphological patterns. By counting the morphs that enter them, and the morphemes included therein and the list between them and the grammatical affixes they share with them at the beginning or at the end of the words [10], [19], [22].

D. Approach based on Graph

The graph approach has been dominant since the 1980s. The finite state approach for morphological analysis was initially studied at Xerox and the first practical application was due to Koskenniemi [20][23]; this has been used to develop wide coverage morphological analyzers for several languages.

In this type of approach, the morphotactic and spelling rules are programmed in a finite state transducer (FST); they require too much manual processing to state rules in an FST and not to analyze words that do not appear in Arabic dictionaries [11][3], [24].

Other analyzers use graphs to perform the morphological analysis of Arabic words [16][17].

E. Statistical Approaches

This approach uses the probability of succession of certain morphemes to perform the morphological analysis of a given word. Statistical data obtained by a corpus allowing to acquire knowledge on the morphology of the language, it learns prefixes, suffixes and patterns from a corpus or a list of words in the target language without any human intervention. This approach uses a list of prefixes, suffixes and patterns to transform from stem to root. The possible prefix-pattern-suffix combinations are constructed for a word in order to obtain the possible roots.

F. Hybrid Approach

Hybrid methods are algorithms that combine several approaches already mentioned. For example in the case of the Buckwalter analyzer, the latter combines linguistic rules (when it introduces the notion of compatibility between prefix, suffix and stem) and dictionaries (when it uses the dictionary of Arabic stems) [14]. Other works use this type of approach [13][15].

III. PRESENTATION OF THE RELATIONAL DATABASE APPROACH

Our new approach is based on the relational concept between tables to perform the morphological analysis of a word.

The database used in this approach uses several tables:
- The tables of proclitics and prefixes of Arabic words.
- The tables of suffixes and enclitics of Arabic words.
- The table of surface patterns of Arabic words.
- The table of Arabic stems.

A. The Tables of Proclitics and Prefixes

The two tables are composed of the Arabic prefixes and the proclitics list. These two tables are related by links which model the prefixes and the proclitics combination rules.

The computability between proclitics and prefixes are going to be explained in Fig. 1.

For example: the prefix “ح” does not combine with the proclitic “ن.”

These Compatibility Tables generates a query (Query1) of prefixes attached to the proclitics which correspond to them, with other information, such as the word type, the pronoun, … (Table I).

B. The Suffix and Enclitic Tables

The two tables are composed of the Arabic suffix and enclitic list. These two tables are related with links which model the rules of combination between suffixes and enclitics.

The computability between suffix and enclitic are going to be explained in Fig. 2.
TABLE I. EXTRACT OF THE REQUEST CREATED BETWEEN THE PREFIXES AND PROCLITICS

<table>
<thead>
<tr>
<th>Prefix + proclitic</th>
<th>Proclitic</th>
<th>Word-type</th>
<th>Gender of a word</th>
<th>Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>فَمَِ</td>
<td></td>
<td>فَمَِ</td>
<td>هو</td>
<td></td>
</tr>
<tr>
<td>انًضارعَ</td>
<td></td>
<td>انًضارعَ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>انًجٓول</td>
<td></td>
<td>انًجٓول</td>
<td></td>
<td></td>
</tr>
<tr>
<td>فعم</td>
<td></td>
<td>فعم</td>
<td></td>
<td></td>
</tr>
<tr>
<td>أَفََ</td>
<td></td>
<td>أَفََ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>أَث</td>
<td></td>
<td>أَث</td>
<td></td>
<td></td>
</tr>
<tr>
<td>فسََ</td>
<td></td>
<td>فسََ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For example: the enclitic "كَ" does not combine with the suffix "تَ".

This database generates a query (Query2) which contains several information, such as word type, gender, pronoun and numbers. To explain more, the following Table II is going to explain the request created between the suffixes and enclitics.

TABLE II. EXTRACT OF THE REQUEST CREATED BETWEEN THE SUFFIXES AND ENCLITICS

<table>
<thead>
<tr>
<th>Suffix + enclitic</th>
<th>Enclitic</th>
<th>Suffix</th>
<th>Pronoun</th>
<th>number</th>
<th>Genre</th>
<th>Type of Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>نحن</td>
<td></td>
<td>نحن</td>
<td></td>
<td>فَمَِ</td>
<td>هو</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. The Tables of Surface Patterns and Stems of Non-derived Words

This table is composed of surface patterns stems and stems of non-derived Arabic names. Each record is composed of several information, such as lemma, stem of the surface patterns or word, type of word, gender, number, class of the surface patterns, etc.

D. Morphological Analysis using the Relational Model

Our approach uses the relation concept used in databases, to connect the query and the tables described previously: Query1, Query 2, and Table III. The resulting query is noted Query-main.

TABLE III. EXTRACT OF THE QUERY CREATED BETWEEN SURFACE PATTERNS AND STEMS OF NON-DERIVED WORDS

<table>
<thead>
<tr>
<th>lemma</th>
<th>Stem-voy</th>
<th>Damir</th>
<th>number</th>
<th>sex</th>
<th>Type of Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>المضارع المجهول</td>
<td>مفرد</td>
<td>مفرد</td>
<td></td>
<td></td>
<td>جعفر</td>
</tr>
<tr>
<td>المضارع المجهول</td>
<td>مفرد</td>
<td>مفرد</td>
<td></td>
<td></td>
<td>جعفر</td>
</tr>
<tr>
<td>اسم جذابة</td>
<td></td>
<td></td>
<td>مفرد</td>
<td></td>
<td>جعفر</td>
</tr>
<tr>
<td>اسم حيوان</td>
<td>مفرد</td>
<td>مفرد</td>
<td></td>
<td></td>
<td>جعفر</td>
</tr>
<tr>
<td>اسم علم</td>
<td>مفرد</td>
<td>مفرد</td>
<td></td>
<td></td>
<td>جعفر</td>
</tr>
</tbody>
</table>

So the Relations between queries and tables are going to be explained in Fig. 3.

Fig. 3. Relations between Queries and Tables.

These query and table are linked by several links modeling the compatibility rules between prefix-proclitic stem and suffix-enclitic stem.

To analyze a given word, the system goes through the following steps:

- Partitioning of the word to be analyzed into a set of proclitics, prefixes, stems, suffixes and enclitics.
- Find all the surface patterns associated with each stem.
- Create a query from this information.
- For example, for the word to analyze "فَدَخَهَث", the pattern of the lemma is "جَعَزََ" associated with the proclitic "فَ" and the suffix "تَ", therefore, the stem calculated from the pattern of the lemma is "دَخَم" (Fig. 4).
- Reconstruction of solution lemmas from surface patterns of lemmas obtained in this query.
- Verification of the set of lemmas obtained, with the basis of the lemmas.

Fig. 4. Request from the Information.
Example:

For the word "أموجه", we find among the solution lemmas "مات - يموت" with the prefix "أ", the enclitic "ه" and the time "فعم مضارع معلوم". But when we check in the database of the lemmas, we find that the verb "مات - يموت" does not admit a complement, which therefore implies that each verb derived from this lemma, does not admit any enclitic. This is why this solution will be excluded.

Fig. 5 illustrates the different layers and process of our analyzer.

IV. EXPERIMENTATION AND RESULTS

A. Implementation

We used with the Java language to build our analyzer morphological. Our application contains three layers including, the presentation layer which processes the data reading part and the result display. The second layer contains the implementation of the algorithm where the rules and filters to apply to analyze a given word as input. The third layer deals with the communication with the MySQL database. The Fig. 6 illustrates the different layers of our analyzer.

B. Application

To see the results, we have developed a demo as a web interface that allows a user to analyze a word entered as input.

Fig. 7 shows the demo page, which contains a button to start, a text area for entering a word to analyze. The text area accepts only Arabic letters.

Example:

For the word "أموجه", our system goes through the steps:

- the possible discretizations are showed on Table IV:
- Creation of the table (noted table-sol) of the surface patterns of stems and stems, result is showed on Table V.
- The creation of a query from the table-sol as shown in Table VI.

<table>
<thead>
<tr>
<th>TABLE IV. POSSIBLE DISCRETIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>proclitic</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>أموجه</td>
</tr>
<tr>
<td>أموجه</td>
</tr>
<tr>
<td>موت</td>
</tr>
<tr>
<td>موت</td>
</tr>
<tr>
<td>موت</td>
</tr>
<tr>
<td>موت</td>
</tr>
</tbody>
</table>

FIG. 5. SCHEMA OF THE MORPHOLOGICAL ANALYSIS PROCESS

FIG. 6. ANALYZER IMPLEMENTATION LAYERS.

FIG. 7. USER INTERFACE OF OUR ANALYZER.
The error rate of our analyzer does not exceed 2%. Most of the found errors are related to insufficient corpus used, which means that our approach is robust against possible false solutions. This robustness is validated by our corpus and the number of criteria used to filter the invalid solutions.

V. CONCLUSION

The approach presented in this article, uses the relational concept relative to databases for making the morphological analysis of Arabic words.

In this approach, the Arabic morphological rules are modeled by links between the different tables used in the main database. The main advantage of this approach is its simplicity of implementation. Moreover, all the variations and the morphological rules are included in the relations between tables. The results obtained are satisfactory and show the importance of the proposed approach.

As future work, we will focus on adding new rules to improve results. Also, we are interested to upgrade our database in order to challenge the performance of our method.

REFERENCES


