Affixal Approach versus Analytical Approach for Off-Line Arabic Decomposable Vocabulary Recognition

Slim KANOUN[2], Fouad SLIMANE[1,2], Hanêne GUESMI[2], Rolf INGOLD[1], Adel M. ALIMI[2], Jean HENNEBERT[1,3]

[1] *DIVA Group, Department of Informatics, University of Fribourg, Fribourg, Switzerland*
[2]*REsearch Group on Intelligent Machines (REGIM), ENIS, University of Sfax, Sfax, Tunisia*
[3] *Business Information System Institute, HES-SO // Wallis, Sierre, Switzerland*

Slim.Kanoun@enis.rnu.tn, Fouad.Slimane@unifr.ch, Rolf.Ingold@unifr.ch, Adel.Alimi@enis.rnu.tn, Jean.Hennebert@hevs.ch

Abstract

In this paper, we propose a comparative study between the affixal approach and the analytical approach for off-line Arabic decomposable word recognition. The analytical approach is based on the modeling of alphabetical letters. The affixal approach is based on the modeling of the linguistic entity namely prefix, infix, suffix and root. The experimental results obtained by these two last approaches are presented on the basis of the printed decomposable word data set in mono-font nature by varying the character sizes. We achieve then our paper by the current improvements of our works concerning the Arabic multi-font, multi-style and multi-size word recognition.

1. Introduction

The existing systems for Arabic word recognition have applied, until now, the traditional approaches which are usually used for the other scripts and based on the modeling of graphic entities resulting from the word image pixel analyses as letters (analytical approach) or pseudo-words (pseudo-analytical approach) or global features extracted from word image (holistic approach). In [8], a good survey is presented. This last survey shows that the analytical approach is the only possible approach for open vocabulary recognition.

Since several years, many researchers specialist in the natural language processing showed that the Arabic vocabulary is composed with two sub-vocabularies: decomposable vocabulary (derived from roots) and indecomposable vocabulary (not derived from roots) [2].

The words of indecomposable vocabulary are constituted by the letters succession. Consequently, their recognition should be made by analytical approach. On the other hand, the words of decomposable vocabulary are composed by the four linguistic segments: a prefix, a suffix, an infix and a root. In addition, the decomposable vocabulary is very rich in linguistic concepts encouraging the proposition of innovative approach more original than an analytical approach. To our knowledge, these concepts are used in the post-processing phase of two systems [1] [9] based on an analytical approach. On the other hand, we note that a neural-linguistic approach is proposed recently in [3] for decomposable word recognition. This approach is based on two transparent neural networks, equipped with linguistic knowledge, and specialized in the recognition of the root – from which the word derives – and the scheme (or template) that the word follows. The word is then automatically reconstituted from its root and scheme. In [5][6], we are proposed and validated, on the printed mono-font nature, a new approach, called affixal approach, based on the linguistic concepts of decomposable vocabulary.

In this paper, we propose a comparative study between the affixal approach and the analytical approach for off-line Arabic decomposable word recognition. The analytical approach is based on the modeling of alphabetical letters. The affixal approach is based on the modeling of the linguistic entity of the decomposable words namely prefix, infix, suffix and root.

In the next sections of our paper, we start with a synthetic presentation of the characteristics of the Arabic script. Then, we present details about the linguistic concepts of Arabic vocabulary. Next, we detail our two recognition systems for decomposable vocabulary: the first system is based on the affixal approach and the second is based on the analytical approach. Finally, we present the comparative study between these two last systems on the basis of the printed decomposable word data set in mono-font nature by varying the character sizes and the current improvements of our works concerning the Arabic multi-font word recognition.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.264

2. Characteristic of Arabic script

The characteristics of Arabic script are synthesized in these following points:

- Arabic script is composed of 28 letters which are mainly consonants.
- Arabic language is represented by a cursive script for printed or handwritten text. It is composed of inter-related consonants written from the right side to the left side.
- Some of these consonants change their shapes according to the place where they occur in the word. Several of them have four shapes: isolated, initial, medial and final.
- Six letters have only two shapes. In fact, they are connected to the letters which precede them but not with those that follow them. These letters are thal, Dal, Ra, waw, Zai and Alif. They present and generate words made up of one or several parts. It is agreed to call them generally PAW (Peace of Arabic Word) or “Pseudo-Word” [8]. A “Pseudo-Word”, thus, corresponds to a chain of one or several related letters.
- More than half of Arabic letters include in their shape dots which can be one, two or three dots. These dots can be above or below the character’s body, but never under and above simultaneously. The presence of these dots in their positions allows us to differentiate between letters that belong to the same family shape.
- Some Arabic characters including “loop” character is called generally occlusion shape which differs from one character to another.

3. Linguistic concepts of Arabic vocabulary

Arabic is a Semitic language. Most of the Arabic vocabulary is made up of words which are derived from roots by insertion of prefixes, infixes and suffixes [2]. These words can be conjugated verbs or specific names for the Arabic language like agent, machine names, lawsuit names, time and place names, preference nouns, analogue adjectives, etc. To cover the totality of the Arabic vocabulary, it is necessary to take into account words, which are not derived from roots. These words can be of various origins such as names of countries, cities, numbers, etc.

As for vocabulary, we have words that are derived from roots and called decomposable vocabulary and we have words that are not derived from roots and named indecomposable vocabulary.

The Arabic decomposable vocabulary is composed by a derivation word system based on a root dictionary, a prefix lexicon, an infix lexicon, a prefix lexicon and the linguistic restrictions. In this framework, we note the following properties:

- The roots are made up of three letters. Each letter can be one of the Arabic Alphabet.
- The prefixes are always located at word beginning. The most common prefixes used in the Arabic texts are about 27. The number of letters making a prefix in a word cannot exceed 3. The prefixes are formed of 6 letters among the 28 letters of the Arabic alphabet.
- The suffixes are always located at word ending. The most common suffixes used in the Arabic texts are about 28. The number of letters making a suffix in a word cannot exceed 3. The suffixes are formed of 7 letters among the 28 letters of the Arabic alphabet.
- The letters constituting infixes take their position between root letters to form radical of word. This position is noted by couple (n, m). For example, the infix "وا" has position (2, 3) between the root letters "فعل" to form a radical "فواعل". In this example, letter "waw" (و) and letter "alif" (ا) are respectively second and third letters of the radical. Five letters of the Arabic alphabet constitute infixes. The number of letters forming an infix in a word cannot exceed 2. About 27 infixes are used commonly in the Arabic texts.

To illustrate better, we present in table 1 the affixal decomposition of some examples of decomposable words.

There are two linguistic restrictions of Arabic decomposable vocabulary: the affixal restriction and the semantic restriction.

The affixal restriction is defined by the fact that the combination between the prefix, suffix and infix of an Arabic decomposable word is valid only if this combination is coherent.

Example:

The root "كتب" (to write) is associated with non-valid affixal combination ("تن ", "ا", "ون") to constitute the word "تنكاتبون" which is not accepted by the Arabic language. However, the word "تتكاتبون" is correct thanks to the association of affixal combination ("تتـ", "ا", "ون") with the same root.

The semantic restriction is defined by the fact that the combination between the prefix, suffix and infix cannot be joined with all roots of the Arabic language since the final combination should have a semantic meaning.

Example:

The application of the valid affixal combination ("ت", "ا", "λ") on root "مرض" gives the word «تمارض" which is accepted by the Arabic language. But the result is totally wrong when the same affixal combination is applied to the root "شرب".

The Arabic decomposable words constitute an open vocabulary. In [2], the authors state that approximately 80 words with current usage can be derived from the same root. To explain, decomposable vocabulary that can be generated, we can start from 1600 roots to estimate a minimum of 120000 words. Consequently, decomposable vocabulary size varies according to the number of roots. For example, in [2], the authors present a dictionary containing 10000 roots.

In the framework of the affixal approach, we constitute a dictionary of 98524 decomposable words starting from 807 roots [5][6]. Part of this dictionary is used to generate a new Arabic printed Text image database called APTI [10].

4. Affixal approach versus analytical approach

4.1 Basic ideas

The analytical approach is always limited to the letter recognition and the concatenation then of the letter hypotheses to build word hypotheses. On the other hand and in the framework of decomposable word recognition, the basic idea of affixal approach is not to recognize the letters but it is to know the linguistic entities: the prefix in the beginning of the word, the suffix at the end of the word, the radical in the middle of the word (the infix and the root). This idea allows establishing a linguistic filtering process of word hypotheses using the affixal and the semantic restrictions. In this sense, we consider that the affixal approach can be reducing considerably the hypotheses exploration space and consequently accelerate and simplify the recognition process. This is justified by the following basic ideas:

- the hypotheses exploration field will be limited to 6 letters for the prefixes recognition in the beginning of the word and to 7 letters for the suffixes recognition at the end of the word instead of the 28 letters of the Arabic alphabet where the recognition is made by the analytical approach.
- the hypotheses exploration field will be limited to 5 letters for the infix recognition instead of the 28 letters of the Arabic alphabets where the recognition is made by the analytical approach.
- the hypotheses exploration field will be limited by the affixal restriction for a prefix, an infix and a suffix hypotheses filtering and by the semantic restriction for a word hypotheses filtering instead of the electronic word dictionary or statistical language models used where the recognition is made by the analytical approach.

For illustrate the basic ideas described above, we present in the following part, our two systems used for decomposable word recognition. The first system is based on the affixal approach and the second system is based on the analytical approach.

4.2 Affixal approach

In this section, we detail and we illustrate the affixal approach through the recognition of the word example: "تتناقلون" ("you transmit" or "you exchange" in English). This word is composed of the prefix "تتـ", the suffix "ون" and the radical "ناقل". This radical is composed by a root "نقل" and an infix "ا". In [5] and [6], there are more details concerning the complete algorithm of affixal approach.

Having a decomposable word image to recognize, we extract the dots (the connected component which do not have an intersection with the base line is considered as dot) and we segment it basing on the vertical projection analysis. The result of this algorithm is the elementary segments. An elementary segment could be a letter or a letter part in the case of over-segmentation. Then, we save coordinates and the position compared to the base line (under or above) for each dot, coordinates and the dimensions for each segment.

After segmentation, we recognize each elementary segment using the 7 invariant moments [4] as features vector and the k nearest neighbor classifier [7] with k equals 1. The training data set used in this classification step covers 28 classes and contains 574 elementary segments in printed nature with Arabic Transparent font and with police size 22, 24 and 28.

For prefix and suffix recognition and based on the fact that a prefix or a suffix could be made up of 1 to 5 elementary segments, we started to concatenate the 5 elementary segments at word beginning and the 5 elementary segments at word ending without exceeding of course the total number of elementary segments of the word. This concatenation tries then to constitute the prefix shape hypotheses among the 9 shapes (after the dots elimination from the 27 prefixes) and the suffix shape hypotheses among the 21 shapes (after the dots elimination from the 28 suffixes).

After the prefix and the suffix shapes constitution, we generate the prefix and the suffix hypotheses by dots association on the basis of its coordinates and its position identified and saved in dots extraction and segmentation step. These two figures show the suggestion of 7 prefix hypotheses ("تتـ", ""تنـ, "نتـ", "ننـ", "نـ", "تـ", "λ") and 3 suffix hypotheses ("ون", "ن", λ). The combination of these prefix and suffix hypotheses generate 21 couple (prefix, suffix) hypotheses.

The affixal restriction is applied to select the coherent couple (prefix, suffix) hypotheses generated by the preceding step. For the word "تتناقلون", having the 21 couple (prefix, suffix) hypotheses, only these 6 hypotheses are coherent: ("ون", "تتـ"), ("ن" , "تتـ"), ("ون" , "تنـ"), ("ن", "تنـ"), ("ون", "تـ"), ("ن", "تـ"). Thus, this filtering eliminates non-valid couples (prefix, suffix) without starting the radical recognition and reduces considerably the recognition exploration space.

After prefix and suffix recognition, we try to recognize the letter shapes which make up the word radical. Thus, for each coherent couple (prefix, suffix) hypothesis, we locate elementary segments of radical in word image by drawing aside those constituting prefix hypotheses at the beginning and those making suffix hypotheses at the end. We try, then, to generate the letter shape hypotheses by elementary segment class's concatenation. Each letter shape hypothesis can be formed by the concatenation of 1 to 4 elementary segments. We generate then the letter hypotheses with dots association on the basis of its coordinates and its positions identified and saved in dots extraction and segmentation step. The combination of letter hypotheses generates the following radical hypotheses corresponding to the couple (prefix, suffix) hypothesis: ("ون", "تـتـ"), ("ون", "تـنـ"): "ناقل", "نافل", "تاقل", "تافل", "ثاقل", "ثافل".

At this level, the radical hypotheses are decomposed into couple (infix, root) hypotheses. We use infix lexicon and root dictionary to retain hypotheses accepted by the Arabic language. At this stage, we constitute triplet (prefix, infix, suffix) and root hypotheses. Then, we filter those whose triplet (prefix, infix, suffix) are coherent. The affixal restriction checking of triplet (prefix, infix, suffix) hypotheses filters two non valid hypotheses. Thus, the two valid word hypotheses which remain are: "تتناقلون" and "تتثاقلون".

In this final stage, we validate final Arabic language acceptance of word hypotheses by validating the semantic association between valid triplet (prefix, infix, suffix) hypotheses and root hypotheses. In our example, the two words "تتناقلون" and "تتثاقلون" are valid.

4.3 Analytical approach

For the decomposable vocabulary recognition by analytical approach, we use the letter recognition engine which we use to recognize the radical of a decomposable word by affixal approach. We employ the following steps:

- recognize the elementary segments of the word,
- concatenate the elementary segments to build hypotheses shapes of letters,
- generate letter hypotheses by dots association to the shape hypotheses of letters,
- generate word hypotheses by concatenation of letter hypotheses,
- verify the existence of each word hypothesis in the dictionary of 98524 decomposable words (see section 3).

4.3 Experimental results

To prove the basic ideas presented in this paper, we carried out the recognition of 450 decomposable word images by the two approaches: analytical approach and affixal approach. This last data set is extracted from the dictionary of 98524 words (see section 3). Our idea is based on the choice of word examples by varying the root, the prefix, the infix and the suffix. We then constituted the images corresponding to these 450 words in printed nature with Arabic Transparent font and with police size 26. This data set is scanned on a 300 - dpi - resolution.

The results show a variation of the word hypotheses number (Table 1) and of the allocated time for recognition process (Table 2) between the affixal approach and the analytical approach. These results are obtained using a Celeron computer having 1.73 mega hertz processor speed and 512 mega bytes random access memory.

Table 1. Word hypotheses number variation according to the recognition approach

Word Hypotheses number	Word images number recognized with the affixal approach	Word images number recognized with the analytical approach
[0..150]	82	1
[150..3650]	368	41
[3650..19956829]	0	408

Table 2. Allocated time variation according to the recognition approach

Allocated time in minutes	Word images number recognized with the affixal approach	Word images number recognized with the analytical approach
1..5	25	224
5..10	425	39
>= 10	0	112
>=60	0	75

The analysis of the results presented in table 1 shows a very significant variation between the hypothesis number put forth by using the affixal approach and the hypothesis number using the analytical approach. It is also worth mentioning that for the 450 recognized decomposable words; there are 408 words which the hypothesis number is between the 3650 and the 19956829 using the analytical approach whereas the hypothesis number does not exceed 3650 for the totality of the words using the affixal approach.

Similarly, the results presented in table 2 show also very significant variations between the allocated time using the affixal approach and the allocated time using the analytical one. Let us note within this framework that for the 450 recognized decomposable words, the allocated time for the totality of the words using the affixal approach does not exceed 10 minutes whereas the allocated time exceeds 10 minutes for 191 words using the analytical approach. Among these 191 words, there are 75 words which have been allocated time equal to or higher than 60 minutes. In addition to that, we should remark that for 425 words, the allocated time using the affixal approach is between 5 and 10 minutes whereas only 39 words have the same allocated time using the analytical approach. To conclude, we can confirm that the affixal approach simplifies the decomposable word recognition. Consequently, it is faster than the analytical approach.

5. Current improvement and future works

Face to the segmentation difficulties of Arabic word image into letters, we start to develop another recognition system based in markovian approach which a first version of system is presented in [11]. This system recognizes an open vocabulary and based on Hidden Markov Models (HMMs). In this system, each Arabic word image are transformed into a sequence of feature vectors computed from a narrow analysis window sliding from right to left. This sequence are the observation input to the HMM. An HMM is then used to model the word where states are associated to characters, sub-characters or directly to their variations. The decoding procedure solves in the same time the recognition of words and the segmentation into character models.

In our future work, we planned to integrate the affixal approach in the recognition system based on markovian approach to develop a marko-affixal approach for multi-font, muti-style and multi-size Arabic word and text recognition. This new approach will be validated using our large APTI (Arabic Printed Text Image) database [10]. This database is synthetically generated using a lexicon of 113’284 Arabic decomposable and indecomposable words, 10 Arabic fonts, 10 font sizes and 4 styles. The database contains more than 45 million of single word images and it is freely available for the scientific community.

6. Conclusion

In this paper, we presented a comparative study between the analytical approach and the affixal approach, for Arabic decomposable vocabulary recognition. Contrary to the analytical approach, the affixal approach shows an acceleration of the time allocated for the recognition process and a diminution of the word hypotheses number. This is justified mainly by the fact that the recognition of prefixes at the beginning of words and suffixes at the end of a word reduces the exploration space of hypotheses to a small subset of letters instead of the Arabic alphabet in the case of the analytical approach. In addition, the hypotheses filtering during recognition process by the linguistic restrictions reduce considerably the space exploration in comparison to word hypotheses filtering by verifying their existence in a dictionary of the language in post-treatment. We described then the current improvement and the future works in the perspective to integrate the affixal approach in the recognition system based on markovian approach to develop a marko-affixal approach for Arabic word and text recognition.

7. References

[1] A. Amin, S. Al-Fedaghi, “Machine recognition of printed Arabic text utilizing natural language morphology”, *International Journal Man-Machine Studies*, 1991, vol. 35, pp. 769-788.

[2] A. Ben Hamadou, “A Compression Technique for Arabic dictionaries: The Affix analysis”, *11th International Conference on Computational Linguistics*, Bonn, Germany, 1986, pp. 286-288.

[3] I. Ben Cheikh, A. Belaïd, A. Kacem, “A Novel Approach for the Recognition of a Wide Arabic Handwritten Word Lexicon”, *ICPR’2008,* Florida, USA.

[4] M. K. Hu, “Visual pattern recognition by moments invariants”, *IRE Trans. Information Theory*, IT-8, February 1962, vol. 8, pp. 179-187.

[5] S. KANOUN, A. ENNAJI, A. M. ALIMI, Y. LECOURTIER, “Linguistic Information Integration on The AABATAS Arabic Text Analysis System”, *IWFHR’2002*, Niagara-on-the-Lake, Ontario, Canada, pp. 389 - 394,.

[6] S. KANOUN, A. M. ALIMI, Y. LECOURTIER, “Affixal Approach for Arabic Decomposable Vocabulary Recognition: A Validation on Printed Word in Only One Font”, *ICDAR’2005*, Seoul, Korea, pp. 1025 - 1029.

[7] H.S. Kim, S. B. Park, “A fast k nearest neighbor finding algorithm based on the ordered partition”, *IEEE Trans. Pattern Anal. And Mach. Intell.,Washinton, DC, USA*, 1986*,* Vol. 8, pp. 761 – 766.

[8] L. M. Lorigo, V. Govindaraju, “Offline Arabic Handwriting Recognition: A Survey”, *IEEE Trans. Pattern Anal. And Mach. Intell.*, 2006, Vol. 28(5), pp. 712-724.

[9] T. Sari, M. Sellami, “MOrpho-LEXical analysis for correcting OCR-genereted arabic words (MOLEX)”, *IWFHR’ 2002,* pp. 461-466.

[10] F. Slimane, R. Ingold, S. Kanoun, M. A. Alimi and J. Hennebert, “A New Arabic Printed Text Image Database and Evaluation Protocols”, *ICDAR’2009*, Barcelona, Spain.

[11] F. Slimane, R. Ingold, M. A. Alimi and J. Hennebert, “Duration Models for Arabic Text Recognition using Hidden Markov Models”. *CIMCA 2008*, Vienne, Austria.