



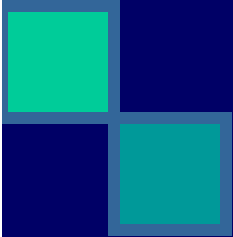

Statistical Methods for Automatic Diacritization of Arabic Text

— ■ ■ ■ —

Moustafa Elshafei¹, Husni Al-Muhtaseb¹, Mansour Alghamdi²
1- King Fahd University of Petroleum and Minerals
2- King Abdulaziz City of Science and Technology



Outline of the presentation



- 
- Introduction
 - Problem Formulation
 - The algorithm
 - The training Set
 - Evaluation of the method
 - Directions for improvement
 - Future work
 - Conclusion
- 



Introduction



Why do we need diacritical Marks?

- 
- Modern Arabic texts are written without the vowel symbols. So, a word such as "علم" can be "عَلِمَ" flag, "عِلْمَ" science, "عُلِمَ" it was known, "عَلِمَ" he knew, "عَلَّمَ" he taught or "عَلَّمَ"
 - Computer applications require Tashkeel, e.g., automatic translation and Arabic text-to-speech.
- 

Introduction (2)

Approaches:

- The Morpho-syntactical approach
(Knowledge based)
- Statistical Pattern Matching Approach
(Data driven approach)

Problem Formulation

			سَمِعَ
			سَمِعَ
			سَمِعَ
حَمَدَ	لَمَّنْ		سَمِعَ
حَمَدَ	لَمَنْ	اللَّهِ	سَمِعَ
حَمَدَ ←	لِمَنْ	اللَّهِ	سَمِعَ
حُمِدَ	لَمَنْ	اللَّهِ	سَمِعَ
حمد	لمن	الله	سمع



Problem Formulation (2)



We need a large set of Vowelized text \mathbf{T}_V ,
corresponding unvowelized text \mathbf{T}_U .

Let $\Gamma(\cdot): \mathbf{L}_V \rightarrow \mathbf{L}_U$

Vocabulary of vowelized words $\mathbf{L}_V = \{v_i\}_1^{N_v}$.

$f_v(k)$ is the frequency of v_k in the training text \mathbf{T}_V

For each word $u_k \in \mathbf{L}_U$, $V_k = \{v \in \mathbf{L}_V; \Gamma(v) = u_k\}$





Problem Formulation (2)

Now, given a word sequence (without diacritical marks)



$$W = w_1 w_2 \dots w_M ; w_t \in \mathbf{L}_U ; t = 1, 2, \dots, M ;$$

determine the most probable diacritized word sequence

$$D = d_1 d_2 \dots d_M$$

Where $d_t = v_j = L_V(j)$

Choose D to maximize the a posteriori probability

$$\hat{D} = \underset{D}{\operatorname{arg\,max}} P(D/W)$$





Problem Formulation (3)



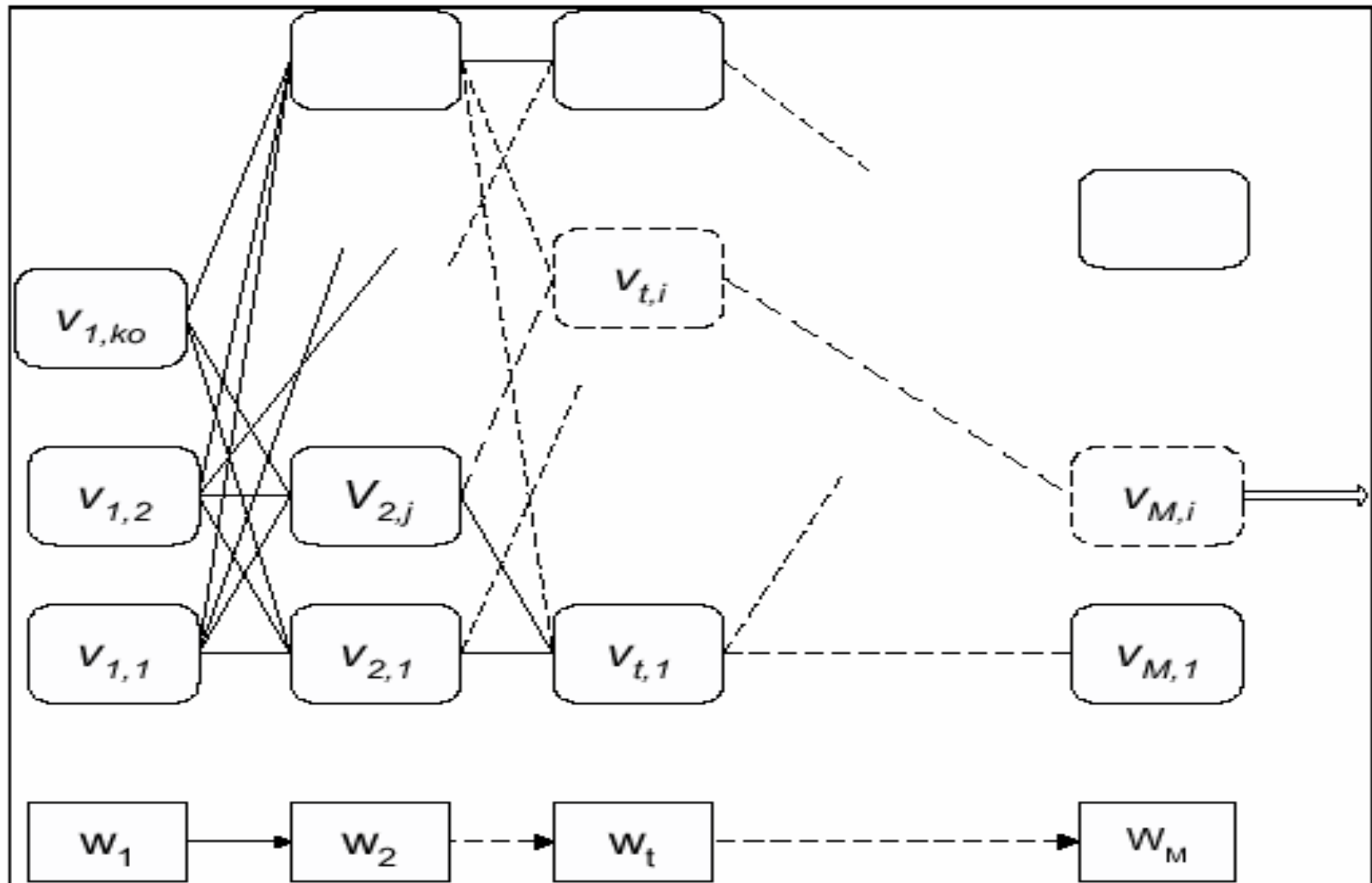
Choose D to maximize the a posteriori probability

$$\hat{D} = \arg \max_D P(D/W)$$


$$P(d_1 d_2 \dots d_m \mid w_1 w_2 \dots w_m) = P(d_1 \mid w_1) \prod_{t=2}^m P(d_t \mid d_{t-1}; w_{t-1} w_t)$$


We need the bigrams $P(v_i \mid v_j) = \frac{C(v_j v_i)}{\sum_{x \in L_v} C(v_j x)} = \frac{C(v_j v_i)}{C(v_j)}$

Viterbi Algorithm



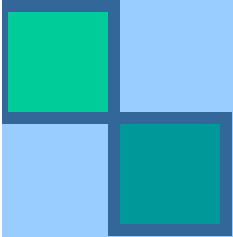


Training Set (Qura'an)

- 78,679 words
 - 607,849 characters with no spacing.
 - Vocabulary : 18,623 diacritized words, consisting of 179,910 characters
 - base words consists of 15,006 words consisting of 80,864 characters.
 - 6807 words appear more than one time
- 



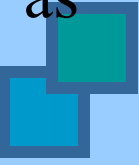
Training Set (2)



Two words $A = a_1 a_2 \dots a_{m_a}$ and $B = b_1 b_2 \dots b_{m_b}$ are considered identical in the regular sense if $R(A, B) = 1$, where $R(A, B)$ is defined as follows

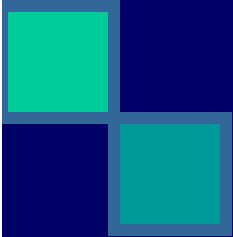

$$R(A, B) = \begin{cases} 1 & \text{if } m_a = m_b \text{ and } a_i = b_i \text{ for } i = 1, 2, \dots, m_a \\ 0 & \text{otherwise} \end{cases}$$

$S(\cdot) : L_v \rightarrow L_v$ which strips words from “Skoon” diacritical marks. Let $S(A) = A^0$, and $S(B) = B^0$. Then two words A and B are said to be R^0 identical, $R^0(A, B) = 1$, if $R(A^0, B^0) = 1$.





Evaluation

- 
- **Phase 1**
 - The test set contains 995 words and 7657 characters. When applying the Viterbi algorithm in its basic form resulted in 230 errors in letter vowelization in 4234 undiacritized characters, that is 5.43% errors in diacritic marks of letters.
 - Analysis of these errors
- 

Evaluation (2)


Problems:

- Fully or partially missing diacritical marks
- Inconsistent representation of *tashkeel* in the training data bases are (لا،لَا، لَا ،) (الَّا،الَّا)
- Tashkeel is based on Quraan recitation (مِّنْ ، مِّن)
- end cases account for 94 errors.
- end-case errors are repeated, and occur in a few frequently used words.
- articles and short words, and accounts for 41 cases. (مِّنْ ، مِّن) ، (إِنَّ ، إِنَّ).



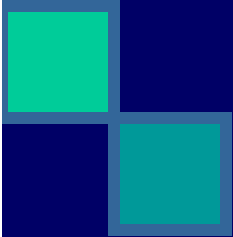

Evaluation

Phase II

- Training using KACST Database of fully Diaritized Text.
 - No End Case
 - The word error rate less than 0.5%.
 - Test set not in the training set: WER 5.5%.
- 

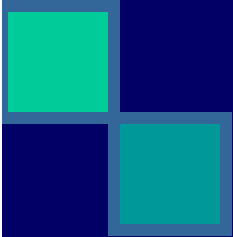


Future Directions

- 
- Words must be in the training set.
 - Use letter/dicritical marks statistics
 - Match morphological patterns
-
- Use trigrams and possibly 4-grams
- 



Conclusion

- 
- The paper proposes the use of Viterbi algorithm to solve the problem of generating the diacritical marks of the Arabic text. The method achieves WER less than 0.5% when tested on sentences from the corpus, and WER of about 5.5% when tested on sentences from outside the corpus.
- 