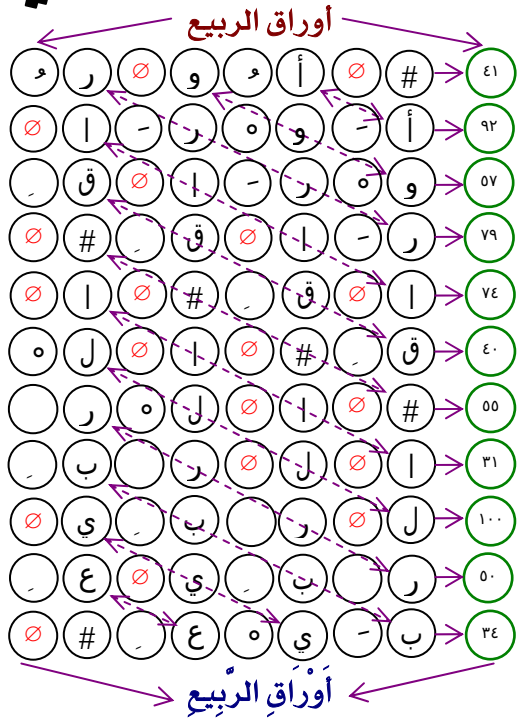
المشكل الآلي



معهد بحوث الحاسب والإلكترونيات

مدينة الملك عبد العزيز للعلوم والتقنية

المملكة العربية السعودية

ح ١٤٢٧هـ

ص ب ٦٠٨٦ الرياض ١١٤٤٢

هاتف: ٤٨١٣٧٦٥ – ١ - ٩٦٦

فاكس: ٤٨١٣٧٦٤ – ١ – ٩٦٦

sstc@kacst.edu.sa

www.kacst.edu.sa

التعريف

تشكل الكتابة العربية المعاصرة اختزالا للكلام قلما نجد مثيلا له في نظم كتابة اللغات الأخرى. فنصف أصوات الكلام تقريبا تحذف رموزها عند الكتابة بالخط العربي. فلا تمثل الصوائت القصيرة ولا التضعيف.

هذا الاختزال في الكتابة يشكل معضلة عند من يتعامل مع النظم الحاسوبية المعاصرة ذات العلاقة بالنص العربي كالناطق الآلي ومحركات البحث والتعرف الآلي على الكلام.

لذا حرص معهد بحوث الحاسب والإلكترونيات على تنفيذ بحوث تقوم بوضع علامات التشكيل على الحروف العربية (التشكيل الآلي). فتم تطوير برنامج "المشكل الآلي" لكي يقوم بتشكيل النص العربي آليا وذلك لاستخدامه في النظم الحاسوبية ذات العلاقة.

المشكل الآلي مستقل بكامل برمجياته وقواعد بياناته وتملك المدينة كامل حقوقه.

المكونات

يتكون نظام المشكل الآلي من التالي:

- خوارزميات لحساب الاحتمالات.
- خوارزميات لاختيار التشكيل ذي الاحتمالية الأعلى.
- ٦٨،٣٧٨ تسلسلا رباعيا للحروف العربية مع تشكيلها واحتمالية التشكيل.

المواصفات

يتسم نظام المشكل الآلي بالأتي:

- نسبة التشكيل الصحيح ٨٧% عل مستوى الحرف بما في ذلك الحرف الأخير من الكلمة.
- سرعة التشكيل = أكثر من ٥٠٠ كلمة في الثانية.
- حجم النظام = ٣ ميغابايت.

المستفيدون

يمكن الاستفادة من المشكل الآلي في:

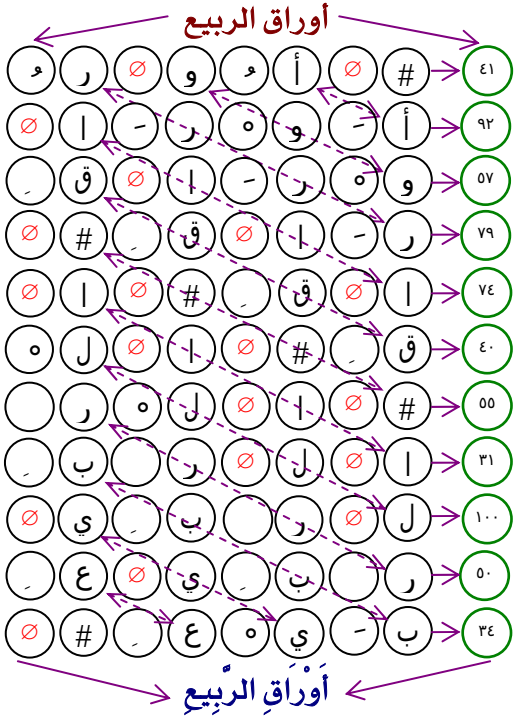
- المراكز البحثية.
- المؤسسات التعليمية.
- مراكز الترجمة.

التطبيقات

يستخدم النظام في الأعمال التالية:

- نظم النطق الآلي.
- نظم التعرف على الكلام.
- محركات البحث.
- نظم الترجمة الآلية.

Arabic Diacritizer



Computer and Electronic Research Institute
King Abdulaziz City for Science and Technology
© 2006

P. O. Box 6086
Riyadh 11442
Kingdom of Saudi Arabia
Telephone: 966-1-4813765
Fax: 966-1-4813764
sstc@kacst.edu.sa
www.kacst.edu.sa

Introduction

Modern Arabic writing includes only the letters that represent the consonants. This means that Arabic vowels and geminates are not represented in the daily writing of Arabic.

The absence of the vocalic and geminate symbols does not allow for full usage of other computational systems such as text-to-speech and automatic speech recognition systems and search engines.

Therefore, CERI has started doing experiments on Automatic Arabic Diacritization to develop a system that can be integrated in other related computer systems. The result is KACST Arabic Diacritizer (KAD).

All the components of KAD are innovated and solely owned by KACST.

Contents

KAD contains the followings:

- Algorithms to calculate the highest probability.
- Algorithms to select the diacritic of the highest probability.
- 68,378 quad-grams of the Arabic letters and diacritics.

Specifications

KAD technical specifications:

- Diacritization accuracy is 87% on the letter level including word-final letters.
- Diacritization speed = more than 500 words/second.
- KAD size = less than 3 MB's.

Beneficiaries

KAD is available at CERI and can be used by:

- Research centers.
- Educational institutes.
- Translation centers.

Applications

KAD can be used in:

- Text-to-speech systems.
- Automatic speech recognition systems.
- Search engines.
- Automatic translation systems.