جامعة السودان للعلوم والتكنولوجيا

كلية الدراسات العليا

أداة مكملة لضبط الكلمات العربية بالشكل

A Complementary Tool for Diacritizing Arabic Words

إعداد:

الروضة عبد اللطيف عبد الحليم حامد

إشراف:

أ.د. عز الدين محمد عثمان

بحث جزئي مقدم لنيل درجة الماجستير في علوم الحاسوب

أغسطس2011

بسم الله الرحمن الرحيم

اقْرَاْ وَرَبُّكَ الأَكْرَمُ ﴿٣﴾الَّذِي عَلَّمَ بِالْقَلَمِ﴿٤﴾

سورة العلق

إهـــــــــــــــــداء

إلى الحب الأول .. أمي

إلى الحاضر الغائب .. أبي

إلى من شاركوني لحظات السعادة والشقاء طول السنين الماضية .. إخواني

إلى كل من علمني حرفاً .. أساتذتي وأستاذاتي

إلى كل من دعمني معنوياً وشجعني على المضي قدماً

إلى زملائي في كلية النيل والذين كانو معي قلباً وقالبا ..

إلى من لاأجد سواهن عندما تضيق بي الدنيا .. صديقاتي

أُهدي هذا العمل

شكر وعرفان

من لايشكر الناس لايشكر الله. فلله الحمد والمنَّة من قبل ومن بعد.

أتقدم بجزيل الشكر إلى كل من:

1. أ.د. عز الدين محمد عثمان: المشرف على البحث وصاحب الفكرة والذي استمتعت جداً بالعمل والنقاش معه.

2. د. هويدا علي عبد القادر أحمد والتي كان لها دور مقدر في إخراج البحث بصورته النهائية.

3. أ. أشواق محمد صالح والتي كان لها عددٌ من الملاحظات التقنية الهامة.

4. د. صديق عمر صديق الأستاذ بقسم اللغة العربية بجامعة الخرطوم ومعهد البروفيسور عبدالله الطيب.

المستخلص

مع تزايد الإهتمام بمعالجة اللغة العربية الطبيعية –أي تصميم برامج تجعل الحاسوب قادراً على فهم نصوصها المكتوبة والمنطوقة والتعامل معها- برزت الحاجة لأنظمة التشكيل الآلي كمكونات أساسية في أنظمة معالجة اللغة العربية كالترجمة الآلية وتوليد الكلام من النصوص المكتوبة والتلخيص الآلي وغيرها، وتخلو غالبية النصوص العربية المعاصرة من علامات التشكيل ولكي تتمكن أنظمة معالجة اللغة العربية من إتمام مهامها بصورة جيدة فهذا يقتضي أن تكون مدخلاتها من النصوص مُشَكَّلة بالكامل وبصورة صحيحة لأنها لاتستطيع لوحدها اكتشاف التشكيل الصحيح على العكس من الإنسان الذي يعتمد في النطق الصحيح للكلمات على ذخيرته اللغوية التي اكتسبها في سنين حياته بل تعتمد في ذلك على أنظمة التشكيل الآلي والتي تستخدم عدة خوارزميات لاختيار الكلمة ذات التشكيل الأنسب. ومع بداية ثمانينيات القرن الماضي ظهرت العديد من أنظمة التشكيل الآلي التي طورتها شركات عربية وعالمية كصخر، RDI وجوجل واستُخدِمَ بعضها كبنيةٍ تحتية لعددٍ من أنظمة معالجة اللغة العربية.

على الرغم من الدقه العالية لهذه الأنظمة فإنها تخطئ في تشكيل بعض الكلمات مما يستدعي وجود نظام مكمل لها لتصحيح هذه الأخطاء.

يساعد النظام المقترح في تشكيل الكلمات غير المُشَكَّلة أو التي شكلت بصورة خاطئة وهو بذلك يتميز على بقية الأنظمة والتي يتم فيها مراجعة وتصحيح أخطاء التشكيل فيها يدوياً.

ويعتمد النظام المقترح على تقنية ربط الجافا بقواعد البيانات وهو نظام مكمل لأنظمة التشكيل الآلي حيث تمثل مخرجاتها مدخلات له فيقوم بتحديد الكلمات التي يشتبه في أن تكون قد شُكِّلَت بصورة خاطئة –إعتمادا على قاعدة بياناته أو مايسمى بالذخيرة اللغوية- ويقدم للمستخدم عدة خيارات للتشكيل ليختار إحداها.

وقد أثبتت الإختبار ات التي أجريت على النظام قدرته على تقليل نسبة الخطأ في تشكيل الكلمات والتي تحصل عند استخدام الانظمة الموجودة بمفردها.

Abstract

With incremental care of the processing of natural Arabic language –i.e. enabling computer to understand its written and spoken texts and dealing with them- the need appears for automatic diacritization systems as basic components of natural Arabic language processing systems such as automatic translation, generating speech from written texts, automatic summarization etc. The majority of Arabic texts are written without diacritic marks; to enable Arabic language processing systems from achieving their tasks in a good manner inputs have to be a truly fully diacritized texts because these systems cannot discover right diacritization by themselves in opposite to human who depends on his linguistic asset which he gained through his life, but they depend on automatic diacritization systems which use several algorithms to select most appropriate dicritization. At the beginning of eighties of the last century several automatic diacritization systems emerged; these systems has been developed by Arabic and international companies such as Sakhr, RDI and Google; some of them was used as an infrastructure for a number of Arabic language processing systems.

Although of high accuracy of these systems but they diacritize some words wrongly, so there is a need for a complementary system to correct those errors.

The proposed system helps in diacritizing wrongly or none diacritized words, this distinguish it from other systems which need manual revision and correction for diacitization errors.

The proposed system depends on Java Database Connectivity technology; it is a complement for automatic diacritization system, so their output acts as its input, then it specifies the words which may be diacritized in wrong manner –

depending on its database or what we call it linguistic asset- and presents multiple diacritization choices for user to select one of them.

Tests on system approved his ability to decrease percentage of diacritization errors which happens when we use existing systems alone.

المصطلحات

السوابق واللواحق.....Affixes

معالجة اللغات الطبيعية.....Natural language Processing (NLP)

الشفرة الأمريكية القياسية لتبادل المعلومات.. American Standard Code for Information Interchange (ASCII)

وضع الشرطات آلياً .....Automatic hyphenation

الترجمة الآلية .....Automatic translation

القواعد غير المقيدة بالسياق.....Context Free Grammar (CFG)

الذخيرة اللغوية/ المتن/ المكنز اللغوي .....Corpus

اللسانيات الحاسوبية.....Computational Linguistics

التشكيل.....Diacritization

علامات التشكيل .....Diacritics

التشفير.....Encoding

وقفات مزمارية.....Glottal stops

التدقيق النحوي .....Grammar Checking

الجرافيمات .....Graphemes

نماذج ماركوف الخفية .....Hidden Markov Models (HMM)

قائمة بالأشكال

قائمة بالجداول

الفهرس