قوالب قواعد اللغة العربية معتمداً على تقنية المكون البرمجي

باسل أحمد باعزيز ، د. خالد عمر ثابت ، د. حسنين محمد البرهمتوشي

كلية الحاسبات وتقنية المعلومات، جامعة الملك عبدالعزيز، جدة – المملكة العربية السعودية

ربيع ثاني ١٤٣٠هـ – مايو ٢٠٠٩م

المستخلص

في ظل الانتشار الواسع لشبكة الانترنت وخاصة في العالم العربي تظل بعض التطبيقات العربية متواضعة ولا ترقى إلى تطلعات المستخدمين وبخاصة التطبيقات المتعلقة بأصل اللغة ، أنظمة استخراج المعلومات *(search engines)* العربية مثل محركات البحث و *(Text to Speech)*، القارئ الآلي للنصوص *(Information Retrieval Systems)* وغيرها من التطبيقات التي تعتمد على الكلمة *(Automatic Translation)*الترجمة الآلية من حيث بناءها وأصلها وتصريفها. إن هذا البحث يقدم نظام متكامل يسمح بمعرفة أصل الكلمة (الجذر) وعمل مدقق إملائي بمعرفة أصل الكلمة وعمل جميع الاشتقاقات اللازمة لها كما يعطي معلومات تفصيلية عن نوع الفعل وحالته من حيث العدد والزمن الصياغة. يوضح هذا البحث أيضاً إمكانية تشكيل الكلمة بحركاتها الكاملة وبدقة عالية عن طريق استخدام تقنية . كما يوفر النظام إمكانية الحصول على جميع التصريفات الممكنة والخاصة بفعل *wildcard* معين. في هذا البحث تم استخدام مجموعة من القوالب التي تمثل الأنواع الرئيسية للفعل الثلاثي وتصريفها بالإضافة إلى عدد من القواعد التي تجعل عملية توليد جميع الكلمات العربية عملية ممكنة.

Arabic Language Template Grammars Component Based Technology

Basil Ahmed Ba-Aziz Dr. Khalid O. Thabit, Prof. Hassanin M. Al-Barhamtoshy

Faculty of Computing and Information Technology , King Abdul Aziz University, Jeddah

Rabi Thani 1430 H - May, 2009 G

Abstract

The present thesis is a unique foundation for Arabic template grammar methodology supporting Natural Language Generation (NLG). Here all words originating from the trilateral verbs are generated with full diacritics and morphological indicators, using the templatic properties of the Arabic language.

A template is defined as a word with the postfix, infixes and prefix characters are written while the characters of the root are missing. The missing characters of the root are presented by digits 1, 2 and 3. Where the digit represents the first, second and third character of the root respectively. This allows the placement of the characters of any other root in place of the digit respectively. Arabic Template Grammar Methodology (ATGM) is a rule that covers the creation of templates and generation of words from root and inserting them in the templates. This results in 28 million diacritic words from 5,321 roots.

Many heuristic methods are involved in dealing with words with weak characters and words with "hamza" because the characters are seen sometimes flipping or disappearing such as (وفى: يفي ، يوفون ، فِ أكل : تؤكل ،). The “hamza” has five shapes (أ ، إ ، ئ ، ء ، ؤ) on computer and every shape has it is own code which is totally a different character based on the diacritic of previous character. Additional prefixes, suffixes for words based on its type (فس, فhave been added like prefixes (هما,كِ ... , نيِ, ... etc). Suffixes for verbs etc).

Finally, this thesis approach is for finding the trilateral Arabic verb root, with all morphological indicators and possible diacritics for any input rooted word, using template masking rather than stemming techniques.