Abdusalam F.A. Nwesri, S.M.M. Tahaghoghi & Falk Scholer
School of Computer Science and Information Technology
RMIT University, Melbourne, Australia.

معالجة النصوص العربية لأجل الفهرسة والاسترجاع

ملخص

تختلف اللغة العربية عن غيرها من اللغات اللاتينية بكثرة مشتقات كلماتها. للبحث في هذه اللغة يجب معالجة كلماتها وإزالة الزوائد لكي يتم فهرسة الصور المختلفة للكلمة الواحدة تحت مرجع واحد. قام العديد من الباحثين في هذا المجال بتطوير أنظمة لإزالة الأحرف الزائدة في الكلمات العربية وأثبتت هذه الأنظمة أن إرجاع الكلمات إلى أصلها يزيد من دقة استرجاع النصوص العربية. معظم هذه الأنظمة يقوم بإزالة الزوائد بصورة غير دقيقة بحيث يتم مقارنة أوائل وأواخر الكلمات بقائمة من هذه الزوائد وإزالتها بدون التأكد من أن الأحرف المزالة هي فعلا أحرف زائدة. في هذه الورقة البحثية نعرض نظامنا الذي نقوم فيه بإزالة الزوائد بصورة صحيحة باستخدام المعاجم العربية والقوانين النحوية. تبرهن النتائج أن إزالة الأحرف الزائدة بصورة صحيحة يزيد من دقة وكم النصوص المسترجعة.

المفردات

إرجاع الكلمات العربية إلى أصولها Arabic stemming – استرجاع النصوص العربية Arabic information retrieval

1. المقدمة

اللغة العربية هي لغة سامية تبنى كلماتها من كلمات ثلاثية - ورباعية أحيانا- تسمى الجذور. يتم ذلك بزيادة أحرف أو ضمائر في مقدمة الجدر أو مؤخرته أو في وسطه وذلك حسب الجنس والعدد والتعريف والحالة الإعرابية. جعلت هذه الخاصية

للكلمة العربية الواحدة صورا متعددة في النصوص العربية، وهذا ما يجعل البحث عنها أمرا صعبا إذ أن البحث عن أي صورة يجب أن يؤدي إلى إيجاد كافة الصور الأخرى في النص العربي. مثلا الكلمات "مدرسة"، "مدرسات"، "مدرسون"، "المدرسون" هي صور متعددة للكلمة "مدرس". ولذا إذا كان البحث عن كلمة "مدرسون" وجب على نظام البحث أن يجد الكلمات المتبقية الأخرى. يتم إنجاز ذلك عادة بإزالة الأحرف الزائدة من الكلمات العربية قبل علمية الفهرسة.

تقوم معظم الأنظمة المتوفرة حاليا بإزالة الزوائد عن طريق المقارنة كما هو متعارف عليه في اللغة الإنجليزية. مثلا يتم إزالة حرف العطف "و" من أوائل الكلمات عن طريق مقارنة الحرف الأول من الكلمة بهذا الحرف وإزالته إذا كان مساويا له. في الكلمتين "وزير"، "وكيل" مثلا نجد أن هذا الأسلوب يؤدي إلى قلب معاني الكلمات العربية إلى كلمات أخرى مخالفة وهذا من شأنه أن يجعل من يبحث عن الوثائق التي تحتوي كلمة "وزير" أن يسترجع الوثائق التي تحتوي على كلمة "زير" والعكس صحيح.

تستخدم اللغة العربية أيضا التشكيل لتمثيل الشكل الإعرابي للكلمة ضمن الجملة. للأسف الشديد عند غياب التشكيل – وهذا حال معظم النصوص العربية حاليا- يصعب اخذ القرار حول معنى الكلمة كما يصعب تحديد ما إذا كانت الأحرف الأولى أو الأخيرة من الكلمة هي أحرف زائدة. مثلا الحرف الأول من الكلمة "وسادة" قد يكون حرف جر أو عطف إذا اعتبرنا أن الكلمة هي "وَسَادة" ويكون حرفا أصليا إذا كانت الكلمة هي "وِسَادة".

في هذه الورقة نقدم أسلوبا جديدا لحذف الزوائد من الكلمة بصورة صحيحة. تعتمد طريقتنا في إزالة اللواحق باستخدام أحد معاجم اللغة العربية وبعض القواعد النحوية للتأكد من أن الأحرف التي يتم إزالتها هي فعلا أحرف زائدة.

سوف نتطرق في الجزء الثاني إلى طرق إزالة الزوائد المتبع في أنظمة استرجاع اللغة العربية ونتطرق إلى بعض الأخطاء التي تقوم بها هذه الأنظمة. في الجزء الثالث سوف نعرض مثالا لنظامين مختلفين يتم استخدامهما في إزالة الأحرف الزائدة من الكلمات العربية. في القسم الرابع نتطرق إلى تفاصيل النظام الجديد وكيف يتم إزالة الأحرف الزائدة بصورة صحيحة باستخدام أحد المعاجم وبعض القوانين النحوية. يتم عرض تقييم النظام الجديد ووصف طريقة التقييم وعرض النتائج في القسم الخامس. نختم هذه الورقة في الجزء السادس بملخص وبعض التعديلات المقترحة للنظام الجديد.

2. فهرسة الكلمات العربية لغرض الاسترجاع

للبحث في النصوص عموما يتم أولا معالجة النصوص من أجل جمع الصور المختلفة للكلمة الواحدة تحت مرجع واحد في فهرس البحث. ويحتوي الفهرس عادة على مرجع للكلمة وعناوين أو أسماء الوثائق التي وردت فيها، عند البحث عن كلمة ما يتم استرجاع عناوين الوثائق المقابلة لتلك الكلمة من الفهرس ومن تم عرضها.

إحدى الطرق المستخدمة لزيادة كفاءة البحث في نظم استرجاع البيانات هي إزالة الأحرف الزائدة (stemming) وإرجاع الكلمات إلى أصولها ليتم فهرسة الصور المختلفة لنفس الكلمة تحت مرجع واحد بدلا من إنشاء مرجع واحد لكل صورة على حدة.

الكلمات العربية لها مشتقات عديدة وذلك لاتصال الضمائر والأحرف بالكلمة مباشرة وهذا يجعل لنفس الاسم أكثر من ستين صورة مختلفة وللفعل صورا أكثر، لهذا فإن إزالة الأحرف الزائدة المتصلة بالكلمة يقلل بصورة كبيرة من حجم الفهرس ويزيد من كفاءة البحث.

يوجد العديد من الأنظمة التي صممت لهذا الغرض وهي تنقسم إلى قسمين أساسيين هما أنظمة إزالة لواحق الكلمات من أول وآخر الكلمة، وأنظمة إرجاع الكلمات إلى جذروها. في الأنظمة الأولى يتم إزالة اللواحق بناءا على قوائم معدة مسبقا يتم مقارنتها بأول وآخر الكلمة، بينما يتم في الأنظمة الثانية حذف هذه اللواحق بنفس الطريقة ثم مطابقة الكلمة المتبقية بالأوزان الصرفية وإرجاع الجدر.

رغم أن معظم قواميس اللغة العربية مفهرسة باستخدام جدر الكلمة، إلا أن الدراسات والأبحاث أثبتت أن هذا النظام غير فعال لفهرسة النصوص العربية لغرض البحث والاسترجاع، وأن النظام الأول أكثر فعالية منه [7,2]. يرجع هذا إلى أن الجدر عام ويشمل العديد من الكلمات الغير متشابهة بل والمختلفة المعنى في

بعض الأحيان. مثلا الكلمات "عمل"، "معمل"، "عميل"، "معمول" لها نفس الجدر "عمل" ولكنها تختلف في المعنى خصوصا عندما يكون النص والمحتوي مختلفين تماما.

1.2. كيف يتم إزالة الأحرف الزائدة

رغم وجود العديد من الأنظمة الخاصة بإزالة الأحرف الزائدة من الكلمات العربية لأجل الفهرسة [3]. ورغم معرفة الزوائد في اللغة العربية من ضمائر وأحرف، إلا انه لا توجد قوائم مشتركة كليا بين هذه الأنظمة، فمنها من يقتصر على إزالة بعض الأحرف ويترك البعض الآخر، ومنها من يقوم بإزالة معظم هذه اللواحق.

تم تحديد اللواحق في الأنظمة السابقة بعد إجراء تجارب باستخدام الموسوعات النصية المتوفرة، ولم يتم التركيز على إزالة لواحق الكلمات بصورة صحيحة لأن السبب الرئيسي وراء هذه التجارب كان الحصول على دقة أعلى بغض النظر عن التركيز على نوع وعدد اللواحق التي يزيلها النظام.

معظم الأنظمة المتوفرة حاليا تقوم بإزالة الأحرف الزائدة عن طريق مقارنة بداية ونهاية الكلمة بمجموعة من اللواحق المحتملة [1،2،4،6،7]. إذا طابقت بداية أو نهاية الكلمة إحدى هذه اللواحق يتم إزالتها بدون النظر إلى الجزء المتبقي من الكلمة. بعض الأنظمة يشترط لإزالة هذه اللواحق أن يتكون الجزء المتبقي من الكلمة من ثلاثة أحرف أو أكثر.

رغم أن مثل هذه الأنظمة يعطي دقة عالية في استرجاع البيانات، إلا أن بعض العمليات الخاصة بمعالجة النصوص العربية قد تتأثر بهذا الأسلوب. فمثلا الترجمة الآلية Machine translation تحتاج إلى وجود كلمات صحيحة بعد إزالة لواحق الكلمات أو أن النتيجة تكون مغايرة تماما إلى معنى الكلمات المترجمة حيث يضطر نظام الترجمة إلى تحويل الجزء المتبقي من الكلمة إلى أصوات اللغة المقابلة وذلك لعدم وجود الجزء المتبقي من الكلمة في القاموس المستخدم في عملية الترجمة. على سبيل المثال، الكلمة "كتابتها" ترجمت إلى "script" عند استخدام محرك الترجمة الخاص بقووقل (Google)،[1] ولكن عند إزالة اللاحقة الأخيرة من الكلمة دون إزالة الحرف "ـتـ" تاركا الكلمة "كتابت" فإن النتيجة هي "Ktapt" وهذا خطأ إذ أن الحرف "تاء" لم يتم إزالته رغم أنه لاحقة زائدة عن الكلمة الأصلية. ويترتب على هذه الأخطاء أيضا أخطاء في عملية التلخيص الآلي Text summarisation والترجمة باستخدام الموسوعات النصية المتوازية Parallel corpora بين اللغات.

3. الأنظمة السابقة

تم في العقد الماضي تطوير العديد من الأنظمة التي تدعم إزالة لواحق الكلمات العربية من اجل فهرستها واستخدامها في عملية البحث واسترجاع النصوص [3]. نتطرق في هذا الجزء إلى اثنين من هذه الأنظمة، الأول يقوم بحذف اللواحق من مقدمة ومؤخرة الكلمة بينما يقوم الثاني بإرجاع الكلمات إلى جذورها باستخدام

[1] http://translate.google.com

الأوزان الصرفية. تم اختيار هذين النظامين من باب التوضيح فقط حتى يتكون لدى القارئ فكرة عن هذه الأنظمة.

1.3. نظام ليه لاركي

قامت ليه لاركي Leah Larkey بتطوير نظام لإرجاع الكلمات إلى أصولها عن طريق إزالة الأحرف الزائدة اطلقت عليه اسم "light10" [7]. قبل البدء في إزالة الزوائد يقوم هذا النظام بإجراء بعض العمليات التمهيدية لخلق نمط موحد بين الكلمات في النص العربي وتتلخص هذه في:

- إزالة الكلمات المتكررة بكثرة في النص Stopwords مثل أدوات العطف والجر والضمائر المنفصلة.
- إزالة حركات الوقف والفصل مثل الفواصل والنقط وعلامات التنصيص وغيرها.
- إزالة الحركات الإعرابية مثل الفتحة والضمة وغيرهما.
- إزالة الأرقام والأحرف التي لا تنتمي للأحرف العربية.
- تحويل الأحرف أ، إ ، آ إلى ا (ألف بدون همزة).
- تحويل الألف المكسورة الأخيرة في الكلمة إلى ياء.
- تحويل التاء المربوطة الأخيرة إلى هاء.

بعد هذه العملية يقوم هذا النظام بإزالة حرف الواو من أوائل الكلمات. ويشترط لإزالة هذا الحرف أن يكون طول الكلمة المتبقية ثلاثة أحرف فما فوق. يتم بعد ذلك إزالة أداة التعريف "الـ" وما لحقها من أحرف الجر وأدوات العطف فيتم إزالة "الـ"،

"والـ"، "فالـ"، "كالـ"، "بالـ" من أوائل الكلمات بشرط أن يكون طول الكلمة المتبقية بعد عملية الإزالة حرفان أو أكثر. بعد ذلك يتم إزالة اللواحق "ها"، "ان"، "ات"، "ون"، "ين"، "يه"، "ية"، "ه"، "ة"، "ي" من أواخر الكلمة على الترتيب السابق على أن يكون طول الكلمة المتبقية بعد عملية الإزالة حرفان أو أكثر.

يتضح من الخطوات السابقة لإزالة اللواحق أن عددا كبيرا من اللواحق العربية مثل الضمائر وأحرف الجر والعطف تم تجاهلها. وهذا من شأنه أن يترك بعض الكلمات في النص دون إرجاعها إلى أصلها. مثلا ينتج عن إزالة اللواحق من الكلمات "كتابه"، "كتابها"، "كتابهم"، "كتابهما"، "فكتابه"، "بكتابه" خمس مراجع في فهرس البحث لنفس الكلمة وذلك بسبب عدم إزالة كافة اللواحق في هذا النظام. كما أن النظام يقع في إزالة أحرف أصلية على أنها لواحق دون التأكد من أن ما تم حذفه هو فعلا أحرف زائدة متصلة بالكلمة.

2.3. نظام شيرين خوجة

يقوم هذا النظام بإزالة اللواحق من بداية ونهاية الكلمة مع مقارنة ما تبقي من الكلمة بمجموعة من الأوزان الصرفية في كل مرة يتم فيها حذف لاحقة في بداية الكلمة أو نهايتها. ويمكن إيجاز الخطوات التي يقوم بها هذا النظام فيما يلي:

- إزالة الحركات الإعرابية من الكلمة.
- إزالة الكلمات المتكررة بكثرة في النص Stopwords مثل أدوات العطف. والجر والضمائر المنفصلة.

- إزالة حركات الوقف والفصل مثل الفواصل والنقط وعلامات التنصيص وكذلك الأرقام.
- إزالة أداة التعريف "الـ" مع ما يتبعها من حروف الجر والعطف من أول الكلمة.
- حذف حرف العطف "و" من أول الكلمة.
- حذف التاء المربوطة من آخر الكلمة.
- حذف أدوات العطف والجر والأحرف المتصلة بالكلمة.

مقارنة ما تبقى من الكلمة مع الأوزان الصرفية ذات نفس الطول. تتم المقارنة بحيث يتم مطابقة الأحرف عدى "ف"، "ع"، "ل" في الوزن الصرفي مع الأحرف الموجودة في نفس الموقع من الكلمة المتبقية. إذا تطابقت الأحرف يتم استخلاص الجدر عن طريق استخلاص الأحرف المناظرة للأحرف الثلاث السابقة من الكلمة. مثلا لاستخلاص الجدر من الكلمة "يقاتلهم" يتم أولا حذف اللواحق "ي"، "هم" من أول وآخر الكلمة. تم يتم مطابقة "قاتل" مع الأوزان الصرفية التي تحتوي على أربعة أحرف. من الواضح أن الوزن الصرفي الوحيد الذي يطابق هذه الكلمة هو الوزن "فاعل" وذلك لمطابقة الحرف المتبقي في الوزن مع الحرف الثاني "ا". في هذه الحالة يتم التوصل إلى أن الجدر هو "قتل". يجدر الإشارة هنا أن النظام يقوم بتحويل الأحرف "أ"، "إ"، "آ" إلى "ا" أثناء عملية المقارنة.

يقوم النظام أخيرا بالتأكد من أن الجدر الذي تم استخلاصه هو جدر صحيح وذلك عن طريق مقارنة الجدر مع قاموس يحتوي على معظم الجذور. إذا وجد الجدر في قاموس النظام يتم إرجاعه، أما إذا لم يوجد يتم إرجاع الكلمة بدون تغيير.

رغم أن خوجة تنبهت إلى مشكلة حذف أو إزالة بعض الأحرف الأصلية من الكلمة على أنها لواحق، إلا أن النظام يقع في هذه المشكلة، حيث يقوم بإزالة حرف الواو الأصلي من الكلمة على انه حرف عطف. كما أن الأسلوب المتبع في التأكد من أن ما تم حذفه هو لاحقة أو حرف زائد عن الكلمة الأصلية - وذلك عن طريق مقارنة الجذر المستخلص من الكلمة بقاموس الجذور - من شأنه أن يمنع حذف اللواحق التي تلحق بالطرف الآخر، إذ أن النتيجة في النهاية هي إما الجذر إن وجد، أو الكلمة الأصلية.

4. النظام الجديد

بناء على ما تقدم، تم بناء نظام لإزالة لواحق الكلمات مع المحافظة على عدم إزالة الأحرف الأصلية للكلمة قدر الإمكان. لكي يتسنى لنا أن نقوم بتحديد صحة الكلمة قبل وبعد إزالة أي لاحقة، كان لابد من استخدام إحدى المعاجم الخاصة باللغة (Lexicon). لدى تم استخدام المعجم المستخدم في برنامج Microsoft office [8]. وهو معجم يستخدم في تصحيح الأخطاء الإملائية عند كتابة النصوص في محررات النصوص بهذا البرنامج.

لأن المعجم يفترض أن يحتوى على جميع المفردات الصحيحة للغة، يمكن استنباط القواعد النحوية للكلمة وتحديد (حالاتها المختلفة عن طريق إضافة اللواحق المحتملة واختبار وجودها في المعجم). وفيما يلي شرح للعمليات المتبعة في إزالة اللواحق:

1.4. توحيد الأخطاء الإملائية في النصوص العربية

كما لاحظنا أن الأنظمة السابقة تقوم بتوحيد نمط الكتابة في النصوص العربية وذلك لاختلاف نمط الكتابة من شخص إلى آخر. وخلاصة ما تقوم به معظم الأنظمة السابقة هو ما يقوم به نظام لاركي من إزالة التشكيل والفواصل والأرقام، وتوحيد نمط كتابة بعض الأحرف مثل الألف واليا والتاء المربوطة. وما أضفناه هو ما يلي:

- تحويل حرفا الواو والهمزة إذا وجدا متواليين في نفس الكلمة إلى الحرف "ؤ"
- تحويل حرفا الياء والهمزة "يء" إذا جاءا متواليين في نفس الكلمة إلى الحرف "ئ"
- تحويل حرفا الألف المكسورة والهمزة "ىء" إذا جاءا متواليين في نفس الكلمة إلى الحرف "ئ"
- إحلال الألفان المتواليان في نفس الكلمة بـ "ا"
- فصل الكلمة إلى كلمتين إذا وجد في وسطها حرف التاء المربوطة مطلقا. فصل الكلمة إلى كلمتين إذا كان في وسطها حرف لا تتغير صورته إذا أتي وسط الكلمة وتبعه "الـ" مباشرة علي أن يقع هذا الحرف بعد الحرف الرابع في الكلمة وأن (يزيد عدد الأحرف المتبقية بداية من موقع هذا الحرف على ثلاثة أحرف).

2.4. إزالة الكلمات المتكررة

يتم حذف الكلمات المتكررة مثل الضمائر وأحرف الجر والعطف أسماء الإشارة والأسماء الموصولة وغيرها من الأحرف التي تتكرر عادة في النص العربي، إذا أن الاحتفاظ بهذه الكلمات وفهرستها ليس له جدوى حيث أنها تكرر في معظم الوثائق العربية. تم حذف نفس الكلمات المستخدمة في نظام لاركي.

3.4. إزالة اللواحق من أول الكلمة

لإزالة أحرف العطف والجر المتصلة من أول الكلمة تم إتباع ما يلي:

يتم إزالة الحرف الأول من الكلمة إذا كان "و"، "ف"، "ك"، "ب"، أو "لـ" وذلك بعد التأكد من أن هذا الحرف هو حرف زائد وذلك بإحدى الطرق الثلاث المذكورة في [10]. ونذكر منها الطريقة RPR والتي يتم فيها حذف هذه الحروف بتكرار الحرف الأول من الكلمة واختبار ما إذا كانت الكلمة الجديدة متواجدة في المعجم. إذا كانت الكلمة موجودة في المعجم يتم المحافظة على الحرف الأول وذلك لقبول الكلمة حرف الجر أو العطف. مثلا لإزالة حرف الواو في كلمة "وزير" يتم تكرار الحرف الأول فتصبح الكلمة "ووزير" ويتم البحث عن هذه الكلمة في المعجم، وتكون النتيجة أن هذه الكلمة موجودة وبذلك يتم المحافظة على الحرف الأول من هذه الكلمة وتكون النتيجة هو كلمة "وزير" كاملة. في المقابل فإن الكلمة "والوالد" ينتج عنها "ووالوالد" وهذه الكلمة لا توجد في معجم اللغة ولهذا يتم إزالة الحرف الأول من هذه الكلمة لتكون النتيجة "الوالد".

هناك استثناء لهذه القاعدة وهو الكلمات التي تبدأ بالحرف "للـ" وذلك لأن هذا الحرف هو عبارة عن حرف اللام مضافا إلى أداة التعريف "لـالـ" وتم حذف حرف الألف لتسهيل النطق. مثلا إذا طبقت القاعدة السابقة على كلمة "للولد" فإن النتيجة تكون "لولد" بدلا من "ولد". ولهذا يتم التعامل مع هذه اللاحقة بإحلال الحرفين الأولين من الكلمة بحرف اللام. إذا نتج عن هذا كلمة موجودة بالمعجم، كانت النتيجة إزالة الحرف الأول فقط إما إذا لم توجد الكلمة الجديدة بالمعجم فإن النتيجة هي نفس الكلمة بدون الحرفين الأولين.

بعد إزالة هذه الأحرف يتم إزالة أداة التعريف "الـ" من أول الكلمة.

4.4. إزالة لواحق الأفعال

الأحرف التي تلحق مقدمة الأفعال هي "س"، "ي"، "ت"، "ن"، "أ". يسبب حذف هذه الأحرف من بداية الكلمات الكثير من الأخطاء إذا لم يتم بعناية. مثلا هناك بعض الأسماء التي تبدأ بمثل هذه الأحرف مثل كلمتا "سيارة" و"ستارة". لإزالة هذه اللواحق بصورة صحيحة نقوم بالتأكد من أن الأحرف الأولي من الكلمة هي فعلا أحرف زائدة وليست أصلية. ولمعرفة ذلك نقوم بتحديد ما إذا كانت الكلمة اسما أم فعلا. ويتم ذلك بإضافة أحد أحرف الجر إلى بداية الكلمة لمعرفة ما إذا كانت الكلمة فعلا، لان الفعل لا يحتمل إضافة أحرف الجر بينما يحتمل الاسم ذلك. تم استخدام حرف "ك" لهذا الغرض وذلك لأن كلا من حرف الواو واللام لا يمكن استخدامهما لأنهما ينتميان إلى فئة أخرى من الأحرف التي يحتمل الفعل إضافتها إليه. مثلا حرف الواو يمكن آن يكون حرف عطف ويقبله الفعل، واللام يمكن أن تكون لام

التعليل ويقبلها الفعل، كما تم استخدام أداة التعريف لمعرفة ما إذا كانت الكلمة اسما لان الأفعال لا تقبل إضافة أداة التعريف إلى أولها.

إذا كانت الكلمة تبدأ بحرف السين كما هو الحال في الأمثلة السابقة، يتم إضافة حرف الكاف إلى الكلمة وكذلك إضافة أداة التعريف إلى الكلمة واختبار وجود إحدى هاتين النسختين في المعجم. إذا وجدت إحداهما لا يتم حذف أي حرف من بداية هذه الكلمة، أما في الحالة الأخرى فيتم حذف حرف السين وما يليه إن كان ضمن الأحرف التي تسبق الفعل. أما إذا كانت الكلمة تبدأ بحرف الألف، إن لم تحتمل الكلمة إضافة حرف الكاف أو أداة التعريف مع بقائها صحيحة، يتم إزالة حرف الألف الأول من الكلمة.

5.4. إزالة اللواحق التي تلحق آخر الكلمة

بعض اللواحق التي تضاف إلى الكلمة لا تسبب مشكلة كبيرة عند حذفها وذلك لعدم كثرة الكلمات التي تنتهي بأحرف أصليه تطابق هذه اللواحق. ولهذا نبدأ بإزالة الضمائر "ها"، "هم"، "هما"، "هن" متى وجدت في آخر الكلمة.

اللاحقة "ان" تلحق المثني وهي تتواجد في العديد من الأسماء كذلك. مثلا الاسم "بستان" ينتهي بهذه اللاحقة وإزالة هذه اللاحقة ينتج عنه الكلمة "بست" وهذا خطأ يجب تداركه. لإزالة مثل هذه اللاحقة وتفادي الوقوع في مثل هذا الخطأ، نقوم بإحلال هذه اللاحقة بـ"ين" وإزالتها فقط إذا كانت الكلمة الجديدة صحيحة.

تقلب التاء المربوطة إلى "ات" لتكوين جمع المؤنث السالم. ولإرجاع الكلمات التي تنتهي بهذه اللاحقة إلى أخواتها يتم تحويل هذه اللاحقة إذا نتج عن إزالتها كلمة توجد في المعجم أو أن إحلالها بالتاء المربوطة ينتج عنه كلمة صحيحة. فيما عدا ذلك يتم الإبقاء عليها واعتبار هذان الحرفان من أصل الكلمة.

يتم إزالة لاحقة جمع المذكر السالم "ون" إذا نتج عن إحلالها بـ "ين" كلمة صحيحة. إذا كان الناتج كلمة غير صحيحة يتم النظر إلى بداية الكلمة وإزالتها إذا بدأت الكلمة بـ "يـ" أو "سيـ" وذلك لاحتمال أن تكون الكلمة فعلا من الأفعال الخمسة. بالنسبة للاحقة "ين" يتم إزالتها فقط إذا كان ناتج إحلالها بـ"ون" أو "ان" كلمة صحيحة.

يتم إزالة اللواحق "وا"، "ية"، "يه"، "ه"، "ة" من آخر الكلمة من دون أي قيد أو شرط.

يتم إزالة الياء التي تأتي في آخر الكلمة إذا نتج عن إزالتها كلمة صحيحة أو نتج عن إحلالها بكل من "ها" و "ه" كلمتان صحيحتان.

عند إزالة بعض اللواحق نحتاج في بعض الأحيان إلى استخدام الصورة الصحيحة من الحرف. مثلا عند حذف "ها" من الكلمة "مدرستها" تنتج الكلمة "مدرست" وهذه الكلمة غير صحيحة إذ أنها لا توجد في معجم اللغة. كذلك قد ينتج عن إزالة بعض الأحرف بقاء الحرفين "ئ" و "ؤ" في آخر الكلمة. في الحالة الأولى يتم إبدال التاء المفتوحة تاء مربوطة وإذا سبقها ألف يتم استبدالهما معا بتاء مربوطة وتكرار علمية حذف اللواحق من جديد لكي يتم حذفها، أما في الحالة الثانية فيتم إبدال الحرفين بهمزة.

وفيما يلي ملخصا للعمليات التي نقوم بها لإزالة الزوائد من الكلمات العربية:

- إزالة التشكيل والأحرف الأخرى كالفواصل وعلامات التنصيص.
- توحيد نمط الكتابة.
- إزالة الكلمات المتكررة.
- إزالة أحرف العطف والجر.
- إزالة اداة التعريف.
- إزالة ملحقات الأفعال من بداية الكلمة.
- إزالة اللواحق من أواخر الكلمة.

تم تسمية هذا النظام بـ "Restrict"

5. التقييم

تم تجربة هذا النظام على الموسوعة النصية المتوفرة من قبل المجمع اللغوي للبيانات (LDC) Linguistic Data Consortium. وهذه الموسوعة هي عبارة عن 383,872 قصة إخبارية نشرت من قبل وكالة الأنباء الفرنسية Agence France Presse (AFP) في الفترة ما بين سنة 1994 وسنة 2000. لهذه الموسوعة عدد 75 استفسارا مع النتائج المتوقعة لها مقسمة إلى 25 استفسارا تم استخدامها في تقييم أنظمة استرجاع البيانات في مؤتمر استرجاع النصوص لسنة Text Retrieval 2001 Conference (TREC)، و50 استفسارا تم استخدامها في سنة 2002.

تم تنفيذ النظام باستخدام هذه الاستفسارات وحساب النتائج لكل من الـ 25 استفسارا الأولى (TREC 2001) والـ 50 استفسارا الثانية (TREC 2002) والاثنان معا (TREC 2002-2001). تم مقارنة النظام مع نظام لاركي كنقطة أساسية لمعرفة مدى التقدم الذي ينجزه استخدام الطرق الجديدة لإزالة الأحرف الزائدة.

تم استخدام محرك البحث LEMUR لإجراء هذه التجارب.[2] وهذا المحرك يدعم فهرسة النصوص المكتوبة باللغة العربية ويحتوي على نظام لاركي لمعالجة النصوص العربية.

تم تقييم النظام باستخدام المقاييس المعيارية المستخدمة في تقييم أنظمة البيانات وهي: الكم Recall والمتوسط العام للدقة Average Precision [9].يتمثل الكم في عدد الوثائق الصحيحة التي يسترجعها النظام مقسوما على العدد الكلي للوثائق الصحيحة الخاصة بالاستفسارات. أما الدقة فتتمثل في حساب عدد الوثائق الصحيحة التي يرجعها كل استفسار مقسوما على العدد الكلي للوثائق الذي يرجعها النظام لذلك الاستفسار. ويتم حساب متوسط الدقة لكافة الاستفسارات في كل مجموعة بجمع دقة كافة الاستفسارات وتقسيم المجموع على عدد الاستفسارات. كما تم تسجيل الدقة التي ينجزها النظام عند استرجاع وثائق (الدقة @ 5) و10 وثائق (الدقة @ 10).

$$\text{الكم} = \frac{\text{عدد الوثائق الصحيحة التي يرجعها النظام}}{\text{العدد الكلي للوثائق الصحيحة الخاصة بالاستفسارات في الموسوعة}}$$

$$\text{الدقة} = \frac{\text{عدد الوثائق الصحيحة التي يرجعها النظام}}{\text{العدد الكلي للوثائق الصحيحة وغير الصحيحة التي يرجعها النظام}}$$

يوضح الجدول رقم 1 النتائج التي تم الحصول عليها من خلال تنفيذ كلا من نظام لاركي (Ligth10) والنظام الجديد (Restrict).

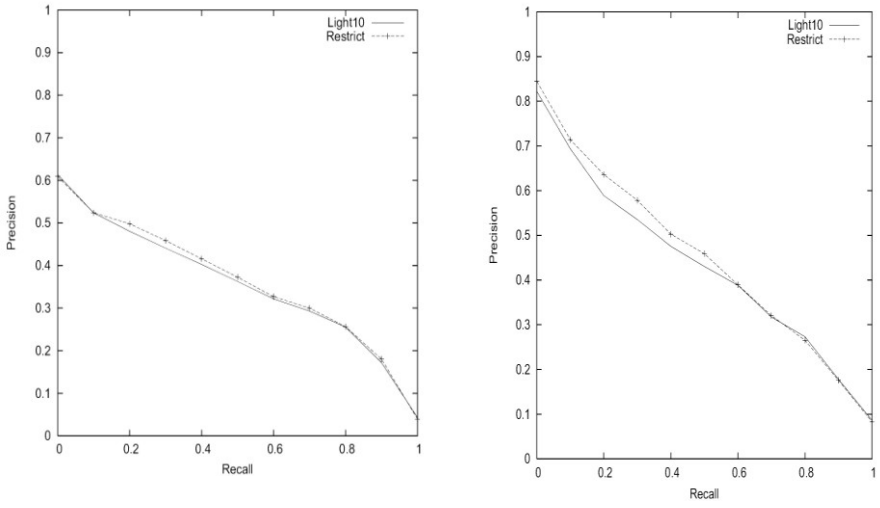
[2]http://www.lemurproject.org

يتضح من هذا الجدول أن النظام الجديد يعطي نتائج أفضل من نظام لاركي الذي يقوم بإزالة اللواحق بدون التأكد من صحة هذه اللواحق. يبين الاختبار الإحصائي *t_test* أن الزيادة التي أضافها النظام الجديد هي زيادة مهمة (*p=0.05372*) عند استخدام كافة الاستفسارات.

من الواضح أن النظام الجديد أضاف زيادة ثابتة في كل من المجموعتين ويتضح ذلك بيانيا في الشكل 1.

TREC 2001-2002 Queries				TREC 2002 Queries				TREC 2001 Queries				
الدقة				الدقة				الدقة				
0@	5@	المتوسط	الكم	10@	5@	المتوسط	الكم	10@	5@	المتوسط	الكم	
506	501	0.372	763	0.440	0.400	0.345	0.836	0.640	0.704	0.425	0.658	Light10
520	520	0.383	784	0.442	0.416	0.352	0.835	0.676	0.728	0.445	0.713	Restrict

جدول رقم 1: نتائج النظام الجديد مقارنة مع نتائج نظام لاركي



استفسارات TREC 2001 **استفسارات TREC 2002**

شكل 1: نتائج تنفيذ النظام على المجموعتين

6. الخلاصة

قمنا في هذه الورقة بعرض نظام جديد يقوم بإزالة الأحرف الزائدة من الكلمات العربية بصورة صحيحة. يختلف هذا النظام عن الأنظمة السابقة في طريقة إزالة الأحرف الزائدة، حيث يقوم بعملية الإزالة فقط إذا تم التأكد من أن الأحرف هي زائدة فعلا عن طريق استخدام أحد المعاجم العربية. تم تحديد العديد من القوانين التي يمكن عن طريقها إزالة هذه الأحرف وتم تجربتها على موسوعة النصوص المعدة لتقييم مثل هذه الأنظمة في سنة 2000، 2001. أثبتت هذه الطريقة أنها فعالة وأنها تزيد من دقة وكم البيانات المسترجعة.

شكر

نتقدم بالشكر الجزيل لأمانة اللجنة الشعبية العامة للتعليم العالي بالجماهيرية العربية الليبية الشعبية الاشتراكية العظمى لدعمها هذا البحث.

المراجع

[2] Aljlayl, M. and Frieder, O.:2002, On Arabic search: improving the retrieval effectiveness via a light stemming approach. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 340–347. ACM Press.

[3] Al-Sughaiyer, I.A. and Al-Kharashi, I. A.:2004, Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

[1] Chen, A. and Gey, F.: 2002, Building an Arabic stemmer for information retrieval. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*. National Institute of Standards and Technology, November, 2002.

[4] Darwish, K. and Oard, D.W.: 2002, Term selection for searching printed Arabic. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 261–268. ACM Press.

[5] Gey, F.C. and Oard. D.W.:2001, The TREC-2001 cross-language information retrieval track: Searching Arabic using English, French or Arabic queries. In *Proceedings of TREC10*, Gaithersburg: NIST.

[6] Khoja, S. and Garside, R.:1999, Stemming Arabic text. Technical report, Computing Department, Lancaster University, Lancaster, September 1999..

[7] Larkey, L.S., Ballesteros, L. and Connell, M. E: 2002, Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 275–282. ACM Press.

[8] Microsoft Corporation. Arabic proofing tools in Office 2003, 2002.

[10] Nwesri, A. F.A., Tahaghoghi, S.M.M. and Scholer, F.: 2005, Stemming Arabic conjunctions and prepositions. In Mariano Consens and Gonzalo Navarro, editors, *Lecture Notes in Computer Science: 3772 - Proceedings of the International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 206–217, Buenos Aires, Argentina, 2–4 November. Springer, Heidelberg, Germany.

[9] Oard, D. W. and Gey, F.C.: 2002, The TREC-2

002 Arabic/English CLIR track. In *TREC*.

URL:http: //www.microsoft.com/middleeast/arabicdev/office/. Office 2003 /Proofing.asp