المحلل الصرفي للغة العربية: المُصَرِّفْ

خالد الحوراني
Khaled@holooli.com

تمهيد:

هذه الورقة تقدم طريقة لمعالجة اللغة العربية صرفياً. المُصرّف يقوم بإيجاد الجذر المحتمل لأي كلمة عربية، ولا يعتمد على قاموس بكلمات اللغة العربية، يتمتع بنسبة عالية من الدقة في إيجاد الجذر والسرعة. أيضاً يتعامل المصرف مع الكثير من الحالات الصعبة والشاذة في اللغة العربية.

تم الاعتماد على اللغة العربية المعاصرة وأخذت المدخلات التجريبية للمصرف من عدة مواقع إخبارية عربية ومواقع الشبكات الاجتماعية للحصول على أكبر قدر من الواقعية.

مقدمة:

تشهد شبكة الإنترنت نمواً هائلاً خاصة بعد ظهور مواقع الويب ٢.٠ (المدونات، المنتديات، إلخ ...). هذا النمو تأثرت به المنطقة العربية فاصبحنا نرى يومياً عشرات المواقع الجديدة.

كان لاختلاف اللهجات في المنطقة العربية أثراً إيجابياً حيث لابد من الكتابة باللغة العربية الفصحى، فالفصحى يفهمها جميع الناطقين بالعربية بخلاف العامية.

ولكن مع تأثر الكثيرين بالعامية وإهمالهم لقواعد اللغة العربية السليمة، ازدادت الأخطاء الإملائية في الكتابة بالفصحى بشكل كبير بحيث أصبح من المستحيل التعامل مع تلك الأخطاء بشكل يدوي ولابد من طريقة آلية لفعل ذلك وهي الفكرة وراء المُصرّف.

معالجة اللغة العربية شهدت تطوراً كبيراً في العقد الأخير واهتماماً متزايداً من مجتمع اللغويين (١، ٣، ٤)، فالكثير من الباحثين العرب والأجانب شرعوا في استكشاف هذا المجال نظراً للتحدي الكبير الموجود فيه، فاللغة العربية تعتبر من أصعب اللغات من الناحية الصرفية (٩).

في اللغة العربية معظم الكلمات يمكن إرجاعها إلى جذرها (مثل: كَتَبَ) ومن هذا الجذر يمكن استخراج عشرات الكلمات (يكتب، كتاب، مكتبة). عملية استخلاص الجذر ليست بالسهلة حيث تتطلب دراية جيدة بقواعد اللغة العربية.

العديد من الأوراق العلمية نشرت في مجال المعالجة الصرفية للغة العربية، وصاحب جزء من هذه الأبحاث تطبيقات لاستخلاص جذر الكلمة (٢، ٣، ٤) وجزء آخر على تطبيقات لاستخلاص جذع الكلمة (كتاب، مكتوب، كاتب) (٨).

أما عن تطبيقات معالجة اللغة العربية فهي كثيرة وهامة، فهي تستعمل في برامج الأرشفة، محركات البحث، الترجمة الآلية، التلخيص الآلي، وبرامج استخلاص المعلومات ومن هنا تأتي أهمية هذا المجال.

معالجة اللغة العربية:

البنية المعقدة للغة العربية تجعل معالجتها أمراً ليس بسيطاً، حيث تتنوع كلمات العربية بشكل كبير وتحتوي على الكثير من الحالات الشاذة التي لاتتبع لقاعدة معينة كجمع التكسير (أقلام). تتنوع الطرق لمعالجة اللغة العربية وكل طريقة لها مزاياها:

١. المقارنة مع قاموس الكلمات: حيث يتم مقارنة الكلمة مباشرة من دون أي تعديل مع قاموس كبير يحتوي على آلاف الكلمات العربية، هذه الطريقة غير فعالة نظراً للحجم الكبير للقاموس (تقدر كلمات اللغة العربية بحوالي ٦×١٠^١٠) وانخفاض الأداء كلما كبر حجم هذا القاموس.
٢. إزالة الزوائد (stemming): الزوائد هي السوابق (ال، بال، و) واللواحق (ات، ون، ها) التي تدخل على الكلمة العربية، والكلمة الناتجة بعد إزالة هذه الزوائد هي الجذع، ويتم تمييز هذه الزوائد من قائمة بتلك الزوائد. من أشهر تطبيقات إزالة الزوائد Al-Stem (٧) وشين (٥). ضعف هذه الطريقة هي ببساطتها حيث لاتستطيع تمميز الحالات الشاذة مثل جمع التكسير وأزمنة الأفعال المختلفة، ولكن تبقى نتيجتها مقبولة.
٣. المعالجة الصرفية: يمكن أن نطلق عليها المرحلة التالية من إزالة الزوائد، حيث تهدف هذه الطريقة للوصول إلى جذر الكلمة وليس جذعها. من أشهر تطبيقات المعالجة الصرفية للغة العربية (٦، ٨).

آلية عمل المُصرّف:

١. **تهيئة النص:** قبل البدء بأي معالجة للنص ينبغي تهيئته كالتالي:
   - إزالة كل شيء عدا الأحرف (أرقام، فواصل، رموز)
   - إزالة التشكيل
   - توحيد شكل حرف الألف باستبدال كل من أ، إ، آ إلى ا
   - تغيير ترميز النص إلى الترميز العالمي الموحد (UTF8)

٢. **إزالة الزوائد:** إزالة الزوائد من بدايات ونهايات الكلمات، في هذه المرحلة قد تحصل بعض الأخطاء مثل إزالة جزء من الكلمة باعتبار أنه زائد (ولد، الطف)، يمكن الحد من هذه الأخطاء بوضع بعض القيود على إزالة الزوائد:

- إزالة "ال" فقط إذا تبقى من الكلمة حرفين أو أكثر، فأقل من حرفين تكون "ال" جزء أساسي من الكلة (الم، الف)
- إزالة "و" لايمكن تقييدها بنفس الشرط السابق حيث يمكن أن تأتي جزءاً من الكلمة بمختلف الأطوال (ولد، وظيفة) أو زائدة (ورب، وقال) لذلك نرسل الكلمتين (مع وبدون "و") إلى المعالج الصرفي.
- استبعاد الكلمات الأجنبية (ألمانيا، كوكتيل) حيث لايجب تمرير هذه الكلمات إلى المعالج الصرفي كونها كلمات غريبة عن العربية لاجذر لها، نقوم مباشرة بمقارنتها بقائمة من الكلمات الأجنبية (كونها لاتخضع لأي قاعدة أو قواعد) للتحقق من صحتها.
- استبعاد الكلمات التي تحتوي على حرفين (أب، هو) لأن المصادر في العربية ثلاثية، رباعية أو خماسية ولايوجد كلمة من حرفين لها مصدر، وكما فعلنا بالكلمات الاجنبية نقارن هذه الكلمات بقائمة للتحقق من صحتها.

٣. **المعالجة الصرفية:** في هذه المرحلة يتم مقارنة الخرج الآتي من المرحلة السابقة مع قائمة من هياكل الجذور كالشكل التالي:

الجذر: كتب

هيكل الجذر: ف X ع X ل (حيث تعبر X عن أي حرف)

الكلمة: كاتب

ولعمل المعالجة الصرفية بشكل سليم ينبغي القيام بعدة عمليات مثل إرجاع الفعل المضارع إلى الماضي وفك التضعيف (قدّم – قددم).

إذا لم نجد تطابقاً بين الكلمة والهياكل الموجودة لدينا فإما الكلمة خاطئة أو تحتاج لمعالجة إضافية فنمررها للمرحلة التالية.

٤. **المعالجة الإضافية:** تصل لهذه المرحلة الكلمات التي تحتاج لمعالجة إضافية كجمع التكسير (مروحة – مراوح) أو الإبدال (وصف – اتصف) أو الإعلال (قاضي – قاضٍ).

الخلاصة:

الطريقتين الأشهر في المعالجة الصرفية للغة العربية إما الوصول لجذر الكلمة أو جذعها، الوصول لجذر الكلمة يتطلب عملاً إضافياً ومن الممكن أن تحدث بهذا العمل الإضافي أخطاء ولكن إذا تمت معالجة هذه الأخطاء تعطي نتيجة أفضل من طريقة الوصول للجذع.

كلتا الطريقتين فعالتين والنتائج تعتبر جيدة ودقتها مقبولة، لكن العمل الإضافي أو المعالجة الإضافية هي مايزيد من دقة أي طريقة، أيضاً أداء وسرعة الطريقة أمر مهم يجب مراعاته جيداً.

المصرّف أعطى نتائج تضاهي النتائج التي تعطيها البرامج الشهيرة، ومع استمرار التطوير به والعمل على تحسينه سيكون له دوراً فعالاً في مجال معالجة اللغة العربية صرفياً بإذن الله.

المصادر:

1. Abu-Salem, H., Al-Omari, M., and Evens, M. Stemming methodologies over individual query words for Arabic information retrieval. JASIS, 50 (6), pp. 524-529, 1999.
2. Al-Fedaghi, S. S. and Al-Anzi, F. S. A new algorithm to generate Arabic root-pattern forms. In Proceedings of the 11th national computer conference. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, pp. 391-400, 1989.
3. Al-Shalabi, R. Design and implementation of an Arabic morphological system to support natural language processing. PhD thesis, Computer Science, Illinois Institute of Technology, Chicago, 1996.
4. Beesley, K. R. Arabic finite-state morphological analysis and generation. In COLING-96: Proceedings of the 16th international conference on computational linguistics, vol. 1, pp. 89-94, 1996.
5. Chen, A. and Gey, F. Building an Arabic stemmer for information retrieval. In TREC 2002. Gaithersburg: NIST, pp 631-639, 2002.
6. Darwish, K. Building a shallow morphological analyzer in one day. ACL 2002 Workshop on Computational Approaches to Semitic languages, pp. 47-54, July 11, 2002.

7. Darwish, K. and Oard, D.W. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. In TREC 2002. Gaithersburg: NIST, pp 703-710, 2002.
8. Khoja, S. and Garside, R. Stemming Arabic text. Computing Department, Lancaster University, Lancaster, 1999. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps
9. McCarthy, John. “Formal Problems in Semitic Phonology and Morphology.” Diss. Indiana University, 1982.