دراسة منهجيات التشكيل الآلي للنصوص العربية بهدف وضع خطة عمل لبناء مشكل آلي مفتوح المصدر

د. ندى غنيم* **د. غيداء ربداوي****

الملخص

يُعدّ غياب التشكيل في النصوص العربية الحديثة من أكبر التحديات التي تواجه المعالجة الآلية للغة العربية. يُمكن للقارئ العربي أن يتوقع التشكيل الصحيح للكلمات عند قراءته نصاً غير مشكول، في حين يحتاج الحاسوب إلى خوارزميات لاستعادة التشكيل بالاعتماد على معارف مختلفة. ونقصد بالتشكيل الحركات جميعها (ضمة، فتحة، كسرة، سكون)، فضلاً عن الشدّة والتنوين. تعتمد بعض منهجيات التشكيل الآلي على المعالجة اللغوية للنصوص، في حين تعتمد منهجيات أخرى على طرائق إحصائية تستند إلى المدوَّنات، وتدمج بعض النظم المنهجيتين السابقتين في مقاربات هجينة.

نعرض في هذا البحث دراسة مرجعية شاملة للطرائق المختلفة التي اعتُمدت في هذه المنهجيات، كما نستعرض المدونات المختلفة التي استُخدمت للاختبارات والتقويم، ثم نطرح مواصفات المدونات التي يجب إعدادها لإجرائيات التقويم، والمعايير التي يجب أن تحققها إجرائية تقويم المشكلات الآلية. يخلص البحث إلى وضع خطة عمل لبناء مشكّل آلي مفتوح المصدر برعاية منظمة ألكسو، وبمشاركة جهات بحثية من بلدان مختلفة.

الكلمات المفتاحية: المعالجة الآلية للغة الطبيعية، التشكيل الآلي للنصوص، مدونات التقويم، تقويم المشكلات الآلية.

*المعهد العالي للعلوم التطبيقية والتكنولوجيا، دمشق-سورية

**المعهد العالي للعلوم التطبيقية والتكنولوجيا، دمشق-سورية.

I. المقدمة:

يُعدّ غياب التشكيل في النصوص العربية الحديثة من أكثر التحديات التي تواجه المعالجة الآلية للغة العربية. يستطيع القارئ العربي أن يتوقع التشكيل الصحيح للكلمات عند قراءته نصاً غير مشكول، أمَّا الحاسوب فلا يُمكنه ذلك مباشرةً بل يحتاج إلى خوارزميات لمحاكاة القدرة البشرية على استعادة التشكيل. ونقصد بالتشكيل الحركات جميعها (ضمة، فتحة، كسرة، سكون)، فضلاً عن الشدّة والتنوين.

تعتمد بعض منهجيات التشكيل الآلي على المعالجة اللغوية للنصوص -أي التحليل الصرفي والنحوي والدلالي والمقامي- في حين تعتمد منهجيات أخرى على طرائق إحصائية تستند إلى المدوَّنات، وتدمج بعض النظم المنهجيتين السابقتين في مقاربات هجينة. قد يعتقد بعضهم أن المنهجيات المعتمدة على المعالجة اللغوية هي الأنجع والأدق، غير أن الأدوات المستخدمة لهذا الغرض في اللغة العربية لا تزال غير ناضجة بما يكفي، ومازالت البحوث جارية لرفع مستوى الدقة في هذه الأدوات، مما أسهم في توجه الباحثين إلى استخدام منهجيات أخرى هجينة نعرض في هذه الورقة مجموعة من البحوث في مجال التشكيل الآلي للغة العربية، ونُفصّل في كل منها الطريقة العامة المتبعة والمدونات المستخدمة ونتائج التقويم، ثم نعرض مجموعة من المعايير المستخدمة في تقويم نظم التشكيل الآلية، وفي الختام نقترح منهجية سيجري اعتمادها في بناء نظام آلي لتقويم المشكّلات الآلية.

II. دراسة مرجعية لنظم التشكيل الآلي

أدت الحاجة إلى تشكيل النصوص العربية آلياً إلى ظهور محاولات عديدة قامت بها شركات مختلفة لتشكيل النصوص العربية بهدف الحصول على مشكّل آلي يُستخدم في نظم حاسوبية أخرى بما في ذلك مركبات الكلام، ومن هذه المحاولات:

- قامت الشركة الهندسية لتطوير نظم الحاسبات RDI ببناء نظام للتشكِّيل الآلي ArabDiac©2.0. اعتمد النظام على أداة التحليل الصرفي ArabMorpho©3.0، وذُكر أن دقته تزيد على 95% مَقيسةً على مستوى الكلمة.
- قامت شركة CIMOS الفرنسية بإنتاج نظام للتشكيل الآلي، وذكر أن نسبة التشكيل الصحيح فيه تساوي 70% تقريباً على مستوى الكلمة.
- قامت "الشركة العالمية" بتطوير نظام تشكيل يقوم بالتشكيل بسرعة عالية ودقة تصل إلى 98% ويسمح بتحديد مستوى التشكيل المطلوب: الأحرف جميعها مع نهايات الكلمات أو دونها. هذا النظام هو جزء من برنامج صخر "أدوات المكتب" office tools، ويعتمد على مستويات مختلفة من معالجة اللغة وتحليلها، بدءاً من الصرف وانتهاءً بفك الغموض في معاني الكلمات. يجري ذلك بتوظيف بحوث معالجة اللغة الطبيعية فضلاً عن قواعد البيانات اللغوية الضخمة التي قامت بتطويرها.
- كذلك قامت شركات أخرى، مثل شركةIBM، وAppTek ، وإنفوأراب، وColtec، بجهود في هذا المجال، ولكن لخدمة برمجيات لها علاقة بالتدقيق الإملائي والنطق الآلي ومعالجة النصوص العربية.

يؤخذ على منتجات هذه الشركات أنها مغلقة المصدر، ولا تُعلن عن خصائصها والأسس التي بُنيت عليها، مما لا يُمكّن المتخصصين من تعديلها لتتلاءم مع نظم وبرمجيات وتطبيقات حاسوبية أخرى. هذا ما دعا عدداً من الباحثين إلى العمل إلى تطوير نظم تشكيل

آلي، وسنقوم في المقاطع التالية بعرض ملخص عن البحوث المنشورة في هذا المجال، ثم سنستعرض المنهجية التي اعتمدت لبناء مشكل آلي مفتوح المصدر.

من الجدير الإشارة، في البداية، إلى أن نظم التشكيل الآلي تختلف تبعاً للموارد التي تعتمد عليها هذه النظم، وعلى الهدف الذي أُعِدَّت من أجله. فبعض النظم تهتم بالتشكيل الكامل للكلمة (مع حركة الحرف الأخير)، وهي تعتمد لذلك على مصادر لغوية مختلفة، وعلى عدد من نظم المعالجة الآلية للغة العربية التي يجب أن تشمل عدة مستويات من المعالجة مستوى إحصائي، ومستوى صرفي، ومستوى نحوي، ومستوى دلالي. ولهذه النظم أهمية بالغة في مجالات استخدام التشكيل الآلي جميعها، وخصوصاً أنظمة تركيب الكلام Text-To-Speech (TTS).

وبالمقابل، هناك نظم تشكيل تهتم بوضع علامات التشكيل على أحرف النص، دون الحرف الأخير من الكلمات. تعتمد هذه الأنظمة على الإحصائيات والمستوى الصرفي، وغالباً لا تتعدى ذلك. ويمكن الإفادة منها في بعض محركات البحث في النصوص المشكولة التي قد لا يؤثر تشكيل أواخر الأحرف تأثيراً كبيراً في استعلامات البحث فيها، فضلاً عن أنظمة تحويل الأسماء المكتوبة بلغة أجنبية إلى اللغة العربية.

يُعتمد في قياس دقة نظم التشكيل عموماً على المؤشرين الآتيين:

- نسبة الخطأ بالكلمات مع الحرف الأخير Word Error Rate (WER). وتعبّر عن نسبة الكلمات المشكّلة خطأً إلى مجموع كلمات النص. وتُعدّ الكلمة مشكّلة خطأً إذا تضمنت خطأً واحداً في التشكيل أو أكثر.
- نسبة الخطأ بالتشكيل Diacritics Error Rate (DER). وتعبّر عن نسبة عدد الأحرف المشكّلة خطأً إلى مجموع أحرف النص.

نعرض فيما يلي دراسة مرجعية لأهم البحوث الجارية في مجال التشكيل الآلي للغة العربية، مبيّنين في كل منها المنهجية المعتمدة والمدونات المستخدمة ونتائج التقويم بدلالة المؤشرات آنفة الذكر.

1- التشكيل باستخدام محوّلات منتهية الحالات مثقلة

نعرض فيما يأتي النظام المعتمد على المحولات منتهية الحالات المثقلة Finite State Transducers [4].

يجري في هذه المنهجية الاعتماد على مدونة نصية مشكولة كلياً لإنشاء نموذج احتمالي مولِّد generative probabilistic model لإجرائية حذف التشكيل، وذلك على شكل متتالية محولات منتهية الحالات. تعمل متتالية المحولات في هذه المرحلة على بناء هذا النموذج الاحتمالي المولّد عن طريق تحويل نص المدونة المشكول تماماً، والمثقّل وفق نموذج اللغة، إلى نصٍ غير مشكول.

يمكن استخدام النموذج الناتج عن مرحلة التدريب لتحديد التشكيل الصحيح للنصوص الجديدة، إذْ يجري استخدام خوارزمية Viterbi [13] على هذا النموذج لإعادة بناء متتالية الكلمات المشكولة ذات الاحتمالية العظمى الموافقة لمتتالية الكلمات غير المشكولة.

يتألف النظام من نموذج 3-gram للكلمات العربية المشكولة. يجري تعلم أوزان هذا النموذج من المدونة النصية المشكولة. لكن، لما كانت اللواصق تشكل تحدياً لنموذج n-gram للكلمات، إذ إنَّ الكلمة نفسها تبدو للنموذج على أنها كلمة أخرى عندما ترد مع لاحق آخر، فقد جرى توسيع النموذج الأساسي بإضافة محلل صرفي بسيط جداً يعمل على فصل اللواصق عن الكلمات. فضلاً عن ذلك، وبهدف تشكيل الكلمات غير

المعروفة للنظام (التي لم ترد في نموذج اللغة الأساسي المعتمد على الكلمات)، جرى تدريب نموذج لغوي 4-gram لأحرف اللغة Letter Language Model (LLM) على متتالية أحرف الكلمات في مجموعة التدريب.

استخدمت للاختبارات مدونة LDC's Arabic Treebark (Part 2) التي تتضمن 501 نص (144.199 كلمة) وأضيف التشكيل ونمط الكلمات POS إلى المدونة. جرى فصل اللواصق عن الكلمات الأصلية، وتوليد شجرة تحليل كاملة بصورة يدوية.

قسمت مجموعة النصوص إلى جزأين: 90 بالمئة للتدريب و10 بالمئة للاختبار. جرى تدريب نموذج للكلمات والأحرف على النسخة المشكولة، ثم طُبِّقَت خوارزمية Viterbi على النسخة غير المشكولة من مجموعة التدريب والمؤلفة من 14000 كلمة. استُخدم نموذج وحيد الكلمة 1-gram دون معالجة اللواصق كقاعدة أساسية، باستخدام تقنية المحولات نفسها. جرى تنفيذ التجربة على مرحلتين: الأولى دون حركات الإعراب والثانية مع الحركات. يبيّن الشكل 1 نتائج الاختبار.

تبيّن النتائج أن استخدام نموذج اللغة ثلاثي الكلمات 3-gram يحسن بشكل قليل نتائج النموذج أحادي الكلمة، في حين أن إضافة نموذج اللواصق، ونموذج اللغة باعتماد الأحرف، يؤدي إلى تحسن ملحوظ من حيث WER و DER.

النموذج	دون الإعراب		مع الإعراب	
	WER	DER	WER	DER
الأساسي	15.48%	17.33%	30.39%	24.03%
3غرام كلمات	14.64%	16.9%	28.42%	23.34%
3غرام كلمات+CC	8.49%	9.32%	24.22%	15.36%
3غرام كلمات+CC+4غرام حروف	7.33%	6.35%	23.61%	12.79%

الشكل 1 نتائج الاختبارات على المدونة

2- نظام حاسوبي لتشكيل النص العربي

نعرض في هذا المقطع مقاربتين قامت بهما مدينة الملك عبد العزيز بالتعاون مع جامعة الملك فهد للبترول والمعادن، وجامعة الملك سعود ووزارة الدفاع والطيران [5].

2-1 المقاربة الأولى: نظام التشكيل باستخدام نموذج ماركوف الخفي Hidden Markov Model (HMM)

أُستخدِم نموذج ماركوف وفق التسلسل الثلاثي للحروف، ولمحاولة رفع كفاءة النظام أضيفت بعض القواعد اللغوية إلى الاحتمالات المتقطعة. جرى اعتماد مدونة تتضمن مجموعة نصوص متنوعة صنفت تحت عدة موضوعات كالرياضة والحوسبة والسياسة والدين والاقتصاد والأدب وغيرها، ثم قام فريق لغوي بتشكيلها. تحتوي هذه الذخيرة على نصوص متوافرة في 231 ملفاً نصياً، كل منها يحتوي على أكثر من 1000 كلمة عربية مشكولة. يبيّن الشكل 2 بعض نتائج الاختبار على هذه المدونة.

الرقم	نصوص التدريب	نصوص الاختبار	نسبة الصحة
١	٥٠٠٠	٥٠٠	٨٣.٠٦
٢	١٠٠٠٠	١٠٠٠	٨٣.١٣
٣	١٥٠٠٠	١٥٠٠	٨٣.٥٣
٤	٢٤٢٨٨	٢٠٠٠	٨٠.٦٩
٥	٢٤٢٨٨	٥٠٠	٧٨.٣٦
٦	٢٤٢٨٨	١٠٠٠	٧٩.٠٣
٧	٢٤٢٨٨	١٥٠٠	٧٩.٦٧
٨	٢٤٢٨٨	٢٤٢٨٨	٨٣.٤١
٩	٥٠٠٠	٣٠٠٠	٨٢.٢٩
١٠	١٠٠٠٠	٣٠٠٠	٨٣.٢٧
١١	١٥٠٠٠	٣٠٠٠	٨٣.٠٣
١٢	٢٠٠٠٠	٣٠٠٠	٨٢.٤٦
١٣	٥٠٠٠	٥٠٠٠	٨٣.٣٥
١٤	١٠٠٠٠	٥٠٠٠	٨٣.٣٥
١٥	١٥٠٠٠	٥٠٠٠	٨٣.١٥

الشكل2 نتائج اختبار نموذج التسلسل الثلاثي للحروف

2-2 المقاربة الثانية: النظام المستقل المعتمد على التسلسل الرباعي للحروف العربية

جرى بناء هذا النظام باعتماد الفرضية القائلة: إنَّه "يمكن تشكيل نص عربي بتمرير احتمالية التشكيل للتسلسل الرباعي للحروف ابتداءً بأول حرف في العبارة وانتهاء بآخر حرف والأخذ بأعلى احتمالية تشكيل لكل حرف". يتكون النظام من وحدتين:

الوحدة الأولى هي وحدة التحليل والمقارنة: ويكون مدخلها النص المطلوب تشكيله وقائمة التسلسلات الرباعية للحروف المشكولة، وتقوم بـ (1) تحليل النص غير المشكول إلى سلسلة من التسلسلات الرباعية للحروف المكونة له، (2) استدعاء التسلسلات الرباعية للحروف المشكولة التي تقابل التسلسلات الرباعية للحروف غير المشكولة. ينتج عن هذه الوحدة قائمة من التسلسلات الرباعية لحروف النص المراد تشكيله مع تشكيلاتها المختلفة واحتمال ورود كل تسلسل. **الوحدة الثانية هي وحدة اختيار التشكيل الأعلى احتمالياً**. يكون خرج الوحدة الأولى مدخلاً للوحدة الثانية، ويكون لكل حرف أربعة احتمالات لتشكيله كحد أعلى، ثم تجمع الاحتمالات وتدمج في احتمال واحد. في النهاية، يجري اختيار التشكيل صاحب الاحتمال الأعلى لكل حرف. لتوضيح هذه الفكرة لنأخذ العبارة "المشكل الآلي للحرف العربي". نفترض أن ناتج وحدة التحليل والمقارنة للسلاسل الرباعية للحروف الأولى من هذه العبارة هو:

#	٠	ا	٠	ل	٠	م	٠	٠,٤٩
ا	٠	ل	٠	م	٠	ش	٠	٠,٣٤
ل	٠	م	٠	ش	٠	ك	٠	٠,٨٩
م	٠	ش	٠	ك	٠	ل	٠	٠,٩٩
ش	٠	ك	٠	ل	٠	#	٠	٠,٤١
ك	٠	ل	٠	#	٠	ا	٠	٠,٣٧

إذْ يعطي العدد في كل سطر احتمال ورود هذه الرباعية في مدونة المقارنة. لتشكيل حرف الشين في هذه العبارة، نجد أن هذا الحرف ورد في التسلسل الثاني والثالث والرابع والخامس، وكان تشكيله في التسلسل الثاني والثالث والرابع" ْ " في حين في التسلسل الخامس كان تشكيله" َ"، هذا يعني أن النظام سيجمع الاحتمالات الواردة في التسلسلات الثلاثة الأولى الموافقة للسكون(0.99+0.89+0.34)، فيكون الاحتمال 2.20 في حين يكون الاحتمال الموافق للفتحة 0.41، وبالنتيجة يأخذ النظام علامة التشكيل ذات المجموع الأعلى وهي هنا السكون.

النتائج: اختُبرت كفاءة النظام على نص افتتاحية جريدة الرياض اليومية - العدد ١٣٨٤٦، وقورنت النتيجة مع خرج نظام التشكيل الذي يعمل بأدوات ماركوف الخفية، فكانت النتيجة ٨٧% :للنظام المستقل و٧٤% للنظام باعتماد نموذج ماركوف.

3- نظام تشكيل باعتماد طرائق تعلم دون إشراف

نعرض في هذا المقطع مقاربة التشكيل التي نفذها Safadi [9]. تعتمد هذه المقاربة على إجراء تحليل صرفي للكلمات باستخدام محلل Buckwalter، الذي يولد (في الحالة العامة) عدة خيارات لكل كلمة، يرتبط كل خيار بتشكيل معين لهذه الكلمة، ومن ثَمَّ تؤول عملية التشكيل الآلي (دون تشكيل الحرف الأخير) إلى انتقاء الخيار الأنسب. يجري لذلك تحديد نوع الكلمة POS من خلال بناء محدد أنماط Tagger للغة العربية. وقد جرى اعتماد طريقة التعلم دون إشراف Unsupervised learning بسبب الافتقار إلى مدوّنة مشكولة باللغة العربية. تعتمد هذه الطريقة في تطبيق الخطوتين السابقتين على كمٍ واسعٍ من النصوص غير المشكولة واستنتاج قواعد إحصائية للتعليم (إحصاء الحالات التي لا لبْس فيها واستنتاج

القواعد منها)، ثم ترتيب هذه القواعد بحسب علامات تأخذ بالحسبان تكرار تطبيقها على النصوص.

لما كانت القواعد المستنتجة آلياً تُغفِل أحياناً بعض الحالات سهلة المعالجة، ولما كانت محدودةً بالقوالب، فقد جرى إضافة قواعد لغوية تجريبية Heuristics (على سبيل المثال: إذا كان طول الكلمة أصغر من ثلاثة، وأحد الاحتمالات الممكنة هو حرف جر PREP اختر هذا الاحتمال).

عند تحديد أنماط نصوص جديدة يجري تطبيق القواعد التي حصلت على أفضل العلامات، ثم القواعد التي تليها بالترتيب بهدف تحسين الأداء (إزالة الغموض أكثر) اعتماداً على الكلمة السابقة ونوعها والكلمة اللاحقة ونوعها.

جرى تنفيذ هذه الطريقة على مدونة من ثلاثة نصوص عربية متنوعة من الموسوعة العربية، يبلغ مجموعها قرابة 18000 كلمة لاستنتاج القواعد. غير أن النتائج لم تقيَّم دقتها لعدم وجود مدونة مشكولة حينها.

4- استنباط التشكيل للنصوص العربية باستخدام منهجية الاعتلاج العظمى

نعرض في هذا المقطع النظام الذي طوره Zitouni [7] لاستنباط تشكيل النصوص العربية بطريقة هجينة اعتمد فيها منهجية الاعتلاج العظمى Maximum Entropy فضلاً عن معالجات لغوية للنصوص المراد تشكيلها، لاستخلاص أنواع مختلفة من المعلومات ومكاملتها منها المفردات والمقطعية وأنماط الكلام.

4-1 السمات المعتمدة

يمكن، في إطار منهجية الاعتلاج العظمى استخدام أي نوع من السمات؛ مما يسمح لمصمم النظام أن يجرب أنواعاً مهمة من السمات بدلاً من أن يهتم بالتفاعلات بين سمة معينة وسمات أخرى. يُمكن تقسيم السمات في النظام إلى ثلاثة أصناف: مفرداتية lexical، مقطعية segment_based، وأنماط الكلام POS. جرى كذلك اعتماد حركات التشكيل الموجودة مسبقاً في النص كسمات إضافية.

- **السمات المفرداتية**: جرى استخدام n-gram حرف بحيث تغطي الحرف الحالي x_i والأحرف التي تسبقه والتي تليه ضمن نافذة طولها 7: $\{x_{i-3},....,x_{i+3}\}$. كذلك، جرى استخدام الكلمة الحالية w_i وكلمات السياق المرافق لها ضمن نافذة طولها 5 (3-gram نحو الأمام ونحو الخلف): $\{w_{i-2},...,w_{i+2}\}$. يجري تحديد هل كان الحرف الذي يجري تحليله واقعاً في بداية الكلمة أو في نهايتها. يضاف إلى ذلك، سمات مشتركة بين مصادر المعلومات المذكورة سابقاً.
- **السمات المقطعية**: تتألف الكلمات المحدّدة بالفراغات من سابق أو أكثر، يليه جذع، ثم لاحق أو أكثر. يسمى الجذع أو كل سابق أو لاحق مقطعاً. يجري عادةً تحديد المعلومات النحوية مثل نمط الكلام أو معلومات التحليل النحوي اعتماداً على المقاطع لا على الكلمات. فمثلاً، تتضمن كلمة "قَابلتهم" الفعل "قَابل"، ولاحقةَ محدد الفاعل المفرد المؤنث الغائب "ت"، ولاحقة الضمير "هم"؛ وهذه الجملة هي جملة كاملة المعنى. لتجزئة الكلمات المفصولة بفراغات إلى مقاطع، يُستخدم نموذج تقطيع تصل دقته إلى 89%. لمحاكاة التطبيقات الحقيقية، تُستخدم المقاطع التي يولدها النموذج بدلاً من المقاطع الفعلية. في نظام التشكيل، يجري تضمين المقطع الحالي a_i، والمقاطع المجاورة ضمن نافذة طولها 5 (3 gram نحو الخلف ونحو الأمام): $\{a_{i-2},...,a_{i+2}\}$. يجري أيضاً تحديد هل كان الحرف الذي يجري تحليله واقعاً في بداية المقطع أو في نهايته؟

- **السمات المتعلقة بنمط الكلمات**: يُربَط مع المقطع a_i المتعلق بالحرف الذي يجري تحليله، نمطُ الكلمة: $POS(a_i)$. يضاف إلى ذلك أيضاً، السمات المشتركة التي تضيف معلومات مفرداتية ومقطعية. يُستخدم نظام تحديد أنماط الكلام الذي يعتمد الطريقة الإحصائية والمبني على معطيات Arabic Treebank [10]. تبلغ دقة هذا النظام نحو 96%.

2-4 المعطيات:

جرى تدريب نظام التشكيل على مدونة Arabic Treebank وتقويمه للأخبار المشكولة Part 3v1.0 ضمن LDC، التي تتضمن 600 وثيقة، فيها 340,281 كلمة. تضمنت مدونة التدريب 288,000 كلمة تقريباً، في حين تضمنت مدونة التقويم 52,000 كلمة.

3-4 **الاختبارات**:

نظراً إلى أنَّ من الشائع كتابة الشدات في النصوص العربية، فقد يكون هدف المشكل الآلي استنباط الحركات (الضمة، الفتحة والكسرة) والتنوين والسكون. جرى تنفيذ دفعتين من الاختبارات: نصوص الاختبار الأول تتضمن الشدّة أساساً، أما نصوص الاختبار الثاني فلا تتضمن أي نوع من أنواع التشكيل. يعمل نظام التشكيل على مرحلتين: الأولى لاستنتاج مواقع الشدات، والثانية للتنبؤ بباقي أنواع التشكيل.

لتقويم أداء النظام، جرى التنفيذ وفق ثلاث حالات بحسب نمط السمات المعتمدة:

1- النظام يعتمد السمات المفرداتية فقط

2- النظام يعتمد السمات المفرداتية والمقطعية.

3- النظام يعتمد السمات المفرداتية والمقطعية وأنماط الكلام.

يُظهر الشكل 3 نتائج تطبيق النظام على نوعي النصوص (مع شدات أو دونها)، ووفق المراحل الثلاث المتعلقة بإضافة السمات المختلفة.

يلاحظ عند استخدام السمات المفرداتية فقط أن نسبة الخطأ في التشكيل على مستوى الأحرف تبلغ 8.2%، ونسبة الخطأ على مستوى الكلمة تبلغ 25.1%، وهي منافسة لقيمة الخطأ في النظام الذي طوره Nelken [4] والذي يستخدم معلومات مفرداتية مقطعية وحرفيَة على مدونة Arabic TreeBank Part2، إذْ بلغت نسبة الخطأ على مستوى الأحرف 12.79%، وعلى مستوى الكلمة 23.61%. تبيّن النتائج أيضاً أنه عند استخدام المعلومات المقطعية تتحسن الدقة إذْ تبلغ نسبة الخطأ على مستوى الكلمة (WER) 18.8% وعلى مستوى المقطع (SER) 9.4% وعلى مستوى الحرف (DER) 5.8%. أمَّا عند إضافة أنماط الكلمات فتتحسن الدقة لتصل نسبة الخطأ على مستوى الكلمة إلى 18%، وعلى مستوى المقطع إلى 8.9%، وعلى مستوى الحرف إلى 5.5%.

الشدة موجودة			الشدة مخمنة		
WER	SER	DER	WER	SER	DER
السمات المفرداتية					
24.8	12.6	7.9	25.1	13.0	8.2
السمات المفرداتية والمقطعية					
18.2	9.0	5.5	18.8	9.4	5.8
السمات المفرداتية والمقطعية وأنماط الكلام					
17.3	8.5	5.1	18.0	8.9	5.5

الشكل 3 نتائج الاختبارات على النصوص مع الشدة أو من دونها

جرى كذلك إجراء اختبارات حيث حذفت حركة تشكيل أواخر الكلمات في معطيات التدريب ولم تجرِ محاولة استنتاجها لاحقاً. يمكن ملاحظة القيم المبيّنة في الشكل 4، وهي نتائج تنفيذ النظام على النصوص السابقة بغض النظر عن حركة تشكيل أواخر الكلمات

التي تتعلق عادة بموقع الكلمات ضمن الجملة.

الشدة مخمنة			الشدة موجودة		
DER	SER	WER	DER	SER	WER
السمات المفرداتية					
3.9	7.0	12.4	3.6	6.6	11.8
السمات المفرداتية والمقطعية					
2.7	4.8	8.6	2.4	4.4	7.8
السمات المفرداتية والمقطعية وأنماط الكلام					
2.5	4.4	7.9	2.2	4.0	7.2

الشكل 4 نتائج الاختبارات دون تشكيل نهاية الكلمات

5- نظام MADA (Morphological Analysis & Disambiguation of Arabic)

نعرض في هذا المقطع نظام التشكيل في اللغة العربية من خلال تحديد أنماط الكلام [1]. يجري في البداية استخدام نظام Buckwalter Arabic Morphological Analysis) BAMA) لتوليد مجموعة التحليلات الممكنة للكلمات. عند إعطاء BAMA كلمات غير مشكولة، فإنه يعيد التحليلات الصرفية الممكنة جميعها، مع التشكيل الكامل لكل تحليل. يجري استخدام نتائج تطبيق مجموعة من محددات أنماط الكلام taggers للانتقاء من بين هذه التحليلات الممكنة. لاختيار أفضل تحليل أنتجه BAMA، تقوم الخوارزمية بعدِّ القيم المخمّنة (المُتنبأ بها Predicted) لمجموعة السمات اللغوية في كل تحليل ممكن. يجري تدريب المصنفات باستخدام SVMTOOL [6] ومجموعة التدريب نفسها المستخدمة في [7] وهي ATB3-Train [3]، وهي جزء من المقطع الثالث من بنك الأشجار العربية لبنسلفانيا Penn Arabic Treebank، المؤلفة من 288000 كلمة. من جهة أخرى، ولإجراء مقارنة بين نتائج MADA والنتائج في [7] فقد جرى استخدام مجموعة واحدة للتطوير والاختبار، وهي ATB3-Devtest، وهي مؤلفة من 52000 كلمة.

عند هذه النقطة يكون نظام MADA-D قد ضيق عدد التحليلات الممكنة (نتائج BAMA) إلى عدد صغير. قد يكون هذا العدد أكبر من الواحد لسببين: الأول أن الطريقة المستخدمة في الاختيار بين التحليلات قد تتوصل إلى أكثر من تحليلٍ بالاحتمال نفسه، والثاني، أنه قد يكون هناك غموض في التشكيل متعلق بالمفردة lexeme-based diacritic ambiguity، ولا يمكن لمحددات الأنماط الصرفية للكلمات أن تزيل غموض التشكيل المفرداتي.

للبت بالغموض المتبقي، جرى تنجيز مكون إضافي كان من الممكن أن يكون نظام إزالة غموض معاني الكلمات WSD، ولكنه أمر صعب، لذلك جرت مقاربة هذا النظام باستخدام نماذج n-gram للغة. جرى استخدام نمطين من المعطيات للتدريب: صيغ الكلمات المشكولة تماماً، ومعطيات جرى فيها الاستعاضة عن كل كلمة مصرّفة inflected بالصيغة المشكولة لسردها citation. للتدريب، استُخدِمت المدونة المستخدمة للتصنيف ATB3-Train، بعد أن جرى تحديد التشكيل وصيغة السرد يدوياً. سميت نماذج اللغة XLM-n، إذْ X يمكن أن تكون "D" في حالة صيغ الكلمات المشكولة تماماً، أو "L" في حالة صيغ السرد، و n هو رتبة الغرام (n=3,2,1). عندما تكون الكلمات المشكولة أو صيغ السرد غير معروفة (ليست ضمن المفردات)، لا يختار نموذج اللغة أياً منها فعلياً. بعد ذلك يجري تطبيق التشكيل وحيد الغرام، وفي النهاية يجري اختيار أحدها عشوائياً.

يُبيّن الشكل 5 نتائج التشكيل ومقارنتها بنتائج Zitouni في [7]، إذْ Only-DLM-1 هي نتائج نموذج وحيد

الغرام مع اختيار عشوائي في حالة الكلمات غير المعروفة، وتظهر أفضل النتائج بالخط الغامق.

النموذج	كل الحركات		إغفال الحركة الأخيرة	
	WER	DER	WER	DER
Only-DLM-1	39.4	14.5	13.8	6.6
tagger-DLM-1	15.9	5.3	6.2	2.5
tagger-DLM-2	15.2	5.1	5.8	2.4
tagger-DLM-3	15.1	5.0	5.7	2.4
tagger-LLM-1	16.0	5.3	6.3	2.6
tagger-LLM-2	15.0	4.9	5.6	2.2
tagger-LLM-3	14.9	4.8	5.5	2.2
Only-LLM-3	35.5	10.8	8.8	3.6
Tagger-noLM	16.0	5.3	6.3	2.6
Zitouni	18.0	5.5	7.9	2.5

الشكل 5 نتائج اختبار نظام MADA

نرى أن نتائج هذا النظام أفضل من نتائج Zitouni، إذْ تقلص الخطأ في حالة التشكيل الكلي للكلمات: WER بـ 17.2% و DER بـ 10.9%، أمَّا في حالة التشكيل عدا نهاية الكلمة فقد تقلص الخطأ بنسبة 30.4% في WER و 12% في DER.

6- تشكيل النصوص العربية كمسألة ترجمة آلية وكمسألة وضع لصاقات لعناصر متتالية

نعرض فيما يأتي نظاماً يعالج التشكيل كمسألة ترجمة آلية رتيبة وكمسألة وضع لصاقات لعناصر متتالية [12].

6-1 التشكيل كالترجمة:

تعدُّ هذه المقاربة النص غير المشكول كنص مصدري والنص المشكول كنص هدف يطلب بناؤه باستخدام الترجمة الآلية الإحصائية المعتمدة على العبارات phrase-based SMT. هذه المقاربة شبيهة إلى حد بعيد بالمنهجية المتبعة في [11].

يجري هنا إنشاء جدولُ عباراتٍ phrase table فيه من جهة المصدر المُدخلاتُ غير المشكولة، ومن جهة الهدف مقابلاتها المشكولة، كذلك يجري بناء نموذج للغة اعتماداً على النصوص المشكولة. وهكذا يكون لكل كلمة أو عبارة مشكولة صيغة واحدة مقابلة لها غير مشكولة. يبنى هذا الجدول، اعتماداً على نص التدريب، بالعبور على الجمل والمواقع جميعها في الجمل، وكتابة كل كلمات n-gram التي تبدأ من كل موقع، إذْ n=1..N. ثم نحصل بحذف تشكيلها على المداخل غير المشكولة ومقابلاتها المشكولة.

اعتُمِدت تمثيلات مختلفة للنصوص المشكولة وغير المشكولة، بهدف تشغيل النظام على مستوى الحرف أو الكلمة أو كلا المستويين. يقوم النظام في الحالة العامة باستخدام المستويين معاً، فإذا لم تكن الكلمة واردة في معطيات التدريب، يُمكن للنظام أن يولد التشكيل بالاعتماد على الأحرف.

6-2 التشكيل كمسألة وضع لصاقات لعناصر متتالية:

يُبيّن المثال الآتي كلمة غير مشكولة والتشكيل الموافق لها مكتوبة باصطلاحات buckwalter.

mwskw	X:	m	w	s	k	w
		|	|	|	|	|
muwsokuw	Y:	u	ϵ	o	u	ϵ

في هذه الطريقة تُمثل الكلمة كمتتالية الأحرف X. في مثالنا: X=(m, w, s, k, w). يجري وضع لصاقة لكل صامت في X تشير إلى تشكيل الحرف في الصيغة Y المشكولة، أمّا الصوامت التي لا نريد تشكيلها فنلصق بها اللصاقة Є. وهكذا تؤول مسألة إيجاد تشكيل X إلى إيجاد المتتالية Y=(u, Є, o, u, Є). وبذا يعدُّ التشكيل كمسألة وضع لصاقات لعناصر متتالية. يستخدم لهذا الغرض طريقة الحقول العشوائية المشروطة Conditional random fields

(CRF) وهي طريقة للنمذجة الإحصائية تقوم بالتنبؤ باللصاقات لعناصر متتالية مع أخذ السياق بالحسبان. استُخدمت في هذا النظام سمات المفردات وأنماط الكلام للكلمة الحالية والسابقة واللاحقة.

3-6 المعطيات:

استُخدمت معطيات مدونة LDC's Arabic Treebark لمقاربتي الترجمة والحقول العشوائية المشروطة، وذلك للتدريب والتعيير والاختبار، وهي المعطيات المستخدمة في [4] و[7].

4-6 النتائج:

1- نظام الترجمة: جرى تقويم النظام باعتماد نسبة الأخطاء في الكلمات WER ونسبة أخطاء التشكيل DER. وقيست النتائج بأخذ تشكيل الحرف الأخير بالحسبان مما يزيد نسبة الأخطاء، ثم دون تشكيله وكانت النتائج على النحو الآتي (الشكل 6):

		مستوى الكلمة	مستوى الحرف
مع تشكيل الحرف الأخير	WER	22.8	21.8
	DER	7.4	4.8
دون تشكيل الحرف الأخير	WER	9.9	7.4
	DER	4.3	1.8

الشكل 6 نتائج اختبار نظام التشكيل (الترجمة)

2-**وضع لصاقات لعناصر متتالية**: لتحسين النتائج السابقة أجري تكامل بين المعلومات القواعدية والإحصائية للتشكيل باعتماد الحقول العشوائية المشروطة (CRF). استخدم نظام Stanford للحصول على نمط الكلام، واستخدم CRF++ toolkit لتدريب النموذج واختباره. ونظراً إلى محدودية الذاكرة فقد جرى التقليل من الخصائص، والاكتفاء بـ 75% من المحتوى مما كان له الأثر في تخفيض الأخطاء. كما تبيّن أن استخدام سياق أطول في النموذج n-gram كان له أثر في تقليص الأخطاء (الشكل 7).

	السياق	2-gram	4-gram	6-gram	8-gram	10-gram	12-gram
مع تشكيل الحرف الأخير	WER	22.8	24.1	22.6	22.2	22.0	21.9
	DER	5.1	5.4	4.9	4.8	4.7	4.7
دون تشكيل الحرف الأخير	WER	9.4	10	8.5	8.3	8.3	8.4
	DER	2.2	2.4	2.0	1.9	1.9	1.9

الشكل 7 نتائج اختبار وضع لصاقات لعناصر متتالية

7- نظام هجين للتشكيل الآلي للغة العربية

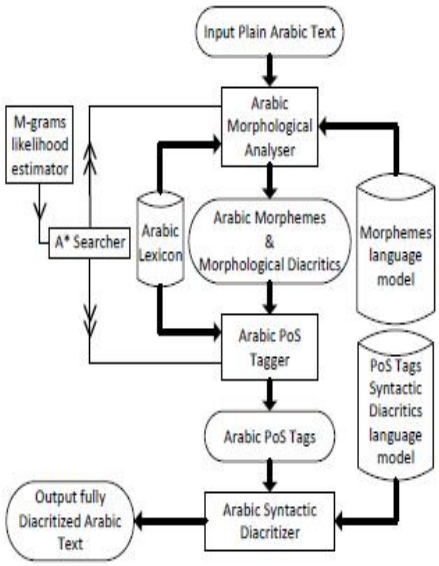
نعرض في هذا المقطع النظام الهجين للتشكيل الآلي للغة العربية [2]. يقوم النظام بتشكيل النصوص العربية باستخدام طبقتين: الأولى هدفها البت بحركات التشكيل الأكثر احتمالاً عن طريق اختيار متتالية تشكيلات الكلمات العربية كاملة الصيغة full-form التي لها أكبر احتمال. عندما تكون الكلمات كاملة الصيغة غير موجودة ضمن المفردات، يجري تطبيق الطبقة الثانية التي تقوم بتحليل الكلمة العربية إلى مكوناتها الصرفية المحتملة (سوابق، جذر، صيغة، ولواحق)، ثم تستخدم التقدير الاحتمالي m-gram للاختيار من بين التحليلات الممكنة للحصول على متتالية التشكيلات الأكثر احتمالاً.

يعتمد هذا النظام على مجموعة تتضمن 62 نمطاً tag تشمل السمات الصرف-نحوية جميعها الخارجة عن السياق للكلمات العربية. جرى استنباط هذه الأنماط من نصوص تتضمن 7800 مفردة محللة مسبقاً، كما

جرى بناء شعاع أنماط كل كلمة (الذي هو دمج لأنماط السوابق والصيغة واللواحق).

1-7 التشكيل عن طريق فك غموض نص محلل إلى عوامل factorized بطريقة إحصائية

يجري في البداية تحليل صرفي للكلمة لاستنباط السوابق والصيغة واللواحق المحتملة، ثم يجري تحديد التحليل الأفضل باستخدام الطرائق الإحصائية لإيجاد متتالية التحليل ذات الاحتمال الأعلى وفق نموذج إحصائي للغة، بُني اعتماداً على مدونة تدريب سبق تحديد أنماطها الصرفية. لتحديد التشكيل النحوي، يجري الحصول على أشعة أنماط متتالية الكلمات العربية مع التشكيل النحوي لكل كلمة، بعد إزالة الغموض الصرفي عنها (الشكل 8). يجري استخدام إزالة الغموض الإحصائي لاستنباط متتالية علامات التشكيل النحوية وأنماط الكلمات التي لها أكبر احتمال وفق نموذج إحصائي للغة بني على مدونة تدريب [8] سبق تحديد أنماط كلماتها وعلامات التشكيل النحوي لكل منها.

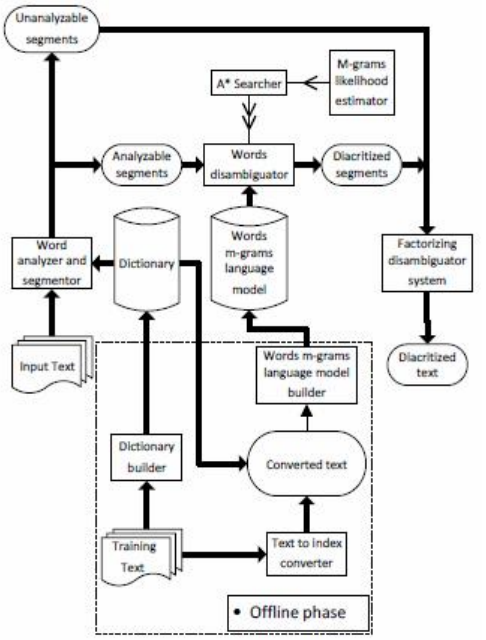


الشكل 8 نظام التشكيل عن طريق فك غموض نص محلل إلى عوامل بطريقة إحصائية

2-7 طريقة فك غموض كلمات هجينة (تامة الصيغة ومحللة إلى عوامل):

لتحسين أداء مشكّل النصوص العربية المحلّلة، جرى تطوير نظام هجين يضم المشكّل المعتمد على الصرف فضلاً عن مشكّل يعتمد على الكلمات تامة الصيغة (الشكل 9)، إذْ جرى استخدام مدونة نصية كبيرة تتضمن التشكيل الصرفي والنحوي لبناء قائمة مفردات عربية تامة الصيغة. في مرحلة أولى، فُهرست هذه المدونة واستخدمت لبناء نموذج إحصائي لغوي للكلمات التامة وفق m-gram.

عند التنفيذ، يجري البحث عن كل كلمة من كلمات النص المدخل ضمن هذا القاموس، فإذا وجدت الكلمة، سميت "قابلة للتحليل"، واستخرجت احتمالات تشكيلها جميعها من القاموس، وهي ما يسمى تحليلات الكلمة. تُسمى مجموعة الكلمات المتتالية القابلة للتحليل "مقطعاً قابلاً للتحليل". تشكل الكلمات المؤلفة للمقطع القابل للتحليل "سلسلة lattice"، يجري إزالة الغموض فيها باستخدام التقدير الاحتمالي وفق m-gram، والبحث عن السلسلة باستخدام خوارزمية A* بهدف استنباط متتالية التشكيل الأكثر احتمالاً. يجري دمج الكلمات التامة التشكيل الموافقة للمقاطع التي جرى فك غموض التحليل فيها مع كلمات الدخل في المقاطع غير القابلة للتحليل، عند وجودها، لتشكيل متتالية كلمات عربية أقل غموضاً. تجري معالجة المتتالية النهائية بعد ذلك بالطريقة المذكورة سابقاً لتحليل المقاطع غير القابلة للتحليل.



الشكل 9التشكيل وفق طريقة فك غموض كلمات هجينة

7-3 المدونات المستخدمة:

I- استخدمت مدونة نصية عربية معيارية TRN-DB-I حجمها نحو 750 ألف كلمة جرى تجميعها من مجالات مختلفة. جرى تحليل هذه المدونة صرفياً، وتحديد أنماط الكلمات فيها وتشكيلها وتدقيقها يدوياً.

II- استخدمت مدونة نصية عربية معيارية TRN-DB-II حجمها 2500 ألف كلمة مشكولة فقط دون أية معلومات أخرى. اِسْتُخْرِجَتْ هذه المدونة من الأدب الإسلامي، ثم حُدِّدَتْ أنماطها يدوياً ولم تُدَقَّقْ سوى مرة واحدة.

III- استخدمت معطيات الاختيار TST-DB وتتضمن 11 ألف كلمة حُدِّدَتْ أنماطها يدوياً من حيث الصرف وأنماط الكلمات والتشكيل. تشمل هذه المدونة مجالات مختلفة، وقد جمعت من مصادر مختلفة عن تلك المستخدمة في I و II.

7-4 الاختبارات:

1- **الاختبار الأول**: تقارن هذه التجربة بين دقة التشكيل الناتجة عن استخدام الطريقتين المعتمدتين على نموذج لغة إحصائي مستنبط من المدونة العربية نفسها. دُرِسَ مدى تغيّر دقة التشكيل لكلتا الطريقتين مع الازدياد المطرد لحجم مدونة التدريب.

يُبيّن الشكل 10 أن نتائج المشكّل الهجين أفضل من المشكّل المحلّل. كما نلاحظ أن الفرق بين نسبتي الخطأ في التشكيل النحوي كبير، في حين أن الفرق بين نسبتي الخطأ الصرفي أقل، ويصغر مع زيادة معطيات التدريب. لذلك، قد تساعد زيادة حجم معطيات التدريب على وصول دقة المشكّل التحليلي إلى المشكّل الهجين، بحيث يمكن التقاط السلوك المعقد للظواهر النحوية فضلاً عن الصرفية.

حجم مدونة التدريب	أخطاء صرفية		أخطاء نحوية	
	مشكل بالتحليل إلى عوامل	مشكل هجين	مشكل بالتحليل إلى عوامل	مشكل هجين
128k	11.5%	9.2%	26.1%	21%
256k	11.8%	7.9%	25.6%	18.7%
512k	9.9%	6.5%	23.3%	16.8%
750k	7.5%	7.0%	24.6%	16.0%

الشكل 10 نتائج الاختبار الأول

2- **الاختبار الثاني**: لما كان الحصول على معطيات تدريب من نمط TRN_DB_II أقل كلفة من نمط TRN_DB_I، فقد أمكن الحصول على مدونة من حجم 2500 ألف كلمة. يسعى الاختبار إلى دراسة تأثير زيادة حجم معطيات التدريب في النموذج الإحصائي غير التحليلي للغة في نسبة أخطاء المشكّل العربي الهجين.

حجم مدونة التدريب	أخطاء صرفية	أخطاء نحوية
حجم TRN_DB_I يساوي K750 كلمة	**7.0%**	**16.0%**
حجم TRN_DB_I + نصف حجم RN_DB_II يساوي k2000 كلمة	**4.9%**	**13.4%**
حجم TRN_DB_I+ حجم RN_DB_II يساوي k3250 كلمة	**3.6%**	**13.0%**

الشكل 11 نتائج الاختبار الثاني

تُبيّن النتائج في الشكل 11 أن دقة التشكيل النحوي قد تصل إلى حد أعلى عندما يتعدى حجم المدونة 2 مليون كلمة. يبدو أنه من الصعب الوصول إلى تحسين واضح عبر الطرائق الإحصائية وحدها بزيادة حجم المدونة فقط، وربما قد يتطلب ذلك دمجها مع أدوات لغوية متطورة.

3- **الاختبار الثالث**: يظهر نسبة أخطاء التشكيل الهجين WER_n وفق مكونتين: التحليلية WER_{fac} وغير التحليلية WER_{unfac}، بحيث $WER_n = WER_{fac} + WER_{unfac}$، يعرض الشكل 12 نتائج هذا الاختبار .

حجم مدونة التدريب	نسبة OOV	أخطاء صرفية		أخطاء نحوية	
		WER_{fac}	WER_h	WER_{fac}	WER_h
TRN_DB_I+ RN_DB_II = k3250 كلمة	**0.1**	**0.0**	**0.0**	**0.1**	**0.1**

الشكل 12 نتائج الاختبار الثالث

8- مقاربة صرفية إحصائية للتشكيل الآلي

نعرض في هذا المقطع النظام المقترح في جامعة محمد الأول للتشكيل الآلي [14] الذي يعتمد منهجية هجينة، إذ تُطبَّقُ طريقة إحصائية على نتيجة المعالجة الصرفية للنصوص. يعمل المشكل الآلي على مرحلتين، يُستخدم في المرحلة الأولى برنامجُ الخليل للتحليل الصرفي الذي يمكّن من الحصول على لائحة الأوزان المشكولة الممكنة لكل كلمة مصحوبة بسابقة ولاحقة الكلمة الخاصين بكل وزن. أمّا بالنسبة إلى الكلمات التي لا وزن لها فتعالج كما يأتي:

- يعوَّض الوزن بالكلمة نفسها في حالة الأدوات (الحروف وأسماء الشرط والأسماء الموصولة وأسماء الإشارة والظروف والضمائر)
- يعوَّض الوزن في حالة الأعلام بكلمة "علم" مع مختلف حالاته الإعرابية.

ونورد في الشكل 13 مثالاً جملة "رحل سالم".

في المرحلة الثانية، يجري الاعتماد على نماذج ماركوف المخفية لتسلسل أوزان الكلمات وخوارزمية Viterbiمن أجل تحديد التشكيل الصحيح للكلمات داخل الجملة.

لم يذكر في المقال أية نتائج تنفيذية، بل ذكر أن العمل جارٍ على بناء مدونة لغوية لاستخدامها في تدريب النظام واختباره.

	السابق	الوزن	اللاحق	مجموع الحالات
رحل	#	فَعَلَ (5 ح.ل)	#	15 حالة
	#	فَعَلَ (5 ح.ل)	#	
	#	فَعَلَ	#	
	#	فَعَلَ	#	
	#	فَعَلَ	#	
	#	فَعَلَ	#	
	#	فَعَلَ	#	
سالم	#	علم (5 ح.ل)	#	12 حالة
	#	فاعل (5 ح.ل)	#	
	#	فَعَلَ	#	
	#	فَعَلَ	#	

الشكل 13 التحليل الصرفي والتعويض لجملة "رحل سالم"

III. المنهجيات العامة المستخدمة في نظم التشكيل الآلي:

تتمحور الطرائق المتبعة جميعها في التشكيل الآلي للنصوص العربية حول إحدى منهجيتين أساسيتين للتشكيل الآلي، أو دمج لهما معاً، وهما:

- المنهجية المعتمدة على السمات اللغوية:

تُعدّ هذه المنهجية الشكل الرئيس والمتعارف عليه من أجل التشكيل الآلي إذْ تتضمن تكاملاً معقداً بين أنظمة معالجة اللغات الطبيعية في مستوياتها المختلفة: الصرفية، والنحوية، والدلالية، والمقامية.

تربط المعالجة الصرفية بين الكلمات غير المشكولة مع النماذج المعروفة للتشكيل (الأوزان الصرفية) ومميزات السوابق واللواحق. وتُسْتَخْدَمُ لهذا الغرض معاجم مشكولة، ومحللات صرفية تدعم التشكيل مثل الخليل الصرفي، ومحلل BAMA، ومحلل Buckwalter.

أمَّا المعالجة النحوية فإنها تحدد تشكيل الحرف الأخير للكلمة من خلال تطبيق محولٍ منتهي الحالات، ويُمكن الإفادة من القواعد النحوية أيضاً في حل مشكلة الغموض في تشكيل كلمة تأخذ

أكثر من نمط كلام (اسم، فعل، ...). ويستخدم لهذا الغرض محددات أنماط الكلام POS taggers ومحللات نحوية.

أمّا فيما يتعلق بالمعالجة الدلالية والمقامية فهما تُستخدمان لإزالة بعض حالات الغموض في معاني الكلمات (مثل كلمة "حسب"، التي يمكن أن تكون "حَسِبَ" -أي ظنَّ- أو "حَسَبَ" -أي عَدَّ-)، ومن ثَمَّ تصفية الحالات المختلفة للتشكيل. ويُستَخدم لهذا الغرض منهجيات لإزالة غموض معاني الكلمات word sense disambiguation تعتمد قواميس أو أنطولوجيات للغة.

- **المنهجية المعتمدة على المعطيات**

تعتمد هذه المنهجية اعتماداً مباشراً على المعطيات، إذْ تستخدم المدوَّنات بهدف استخراج الإحصائيات اللغوية لاستنتاج التشكيل في النص أو لتدريب شبكات عصبونية تسمح بوضع نموذج يمثّل طريقة للتشكيل في اللغة، كما تستخدم هذه الإحصائيات في خوارزميات وطرائق إحصائية- كطريقة نموذج ماركوف المخفي Hidden Markov Model- للحصول على تشكيل كامل للنص العربي.

وبالعودة إلى جميع الطرائق المتبعة في البحوث المذكورة، نجد أنها جميعها طرائق هجينة، إذ إنَّها تدمج ما بين الاعتماد على الموارد اللغوية (وأبسطها المدونات المشكولة والمحللات الصرفية ومحددات أنماط الكلام)، والطرائق الإحصائية المختلفة، مما يبرّر أهمية بناء موارد لغوية ضرورية لبناء المشكلات الآلية ذات الدقة العالية.

IV. معايير تقويم المشكِّلات الآلية

استعرضنا في نظم التشكيل الآلي المختلفة عدداً من المدونات المستخدمة، بعضها مدفوع والآخر متاح مجاناً. تُعدُّ المدونات مكوّناً جوهرياً، لذا فعند دراسة معايير تقويم المشكّلات الآلية لا بدّ من دراسة أسس تقويم المدونات التي تستند إليها اختبارات نظم التشكيل، فضلاً عن معايير تقويم النظم نفسها.

1- مواصفات المدونات العربية الضرورية للتشكيل الآلي

لما كانت المدوَّنات والمعاجم العناصر الرئيسة التي تعتمد عليها نظم التشكيل الآلي، كان لا بُدَّ من وضع خصائص لها. بالرجوع إلى المدونات السابقة وإلى حاجات نظم التشكيل يمكننا أن نلخص الشروط التي يجب أن تحققها المدوَّنة فيما يأتي:

- أن تشمل مجموعة كبيرة من النصوص في مجالات متعددة وتمتد على مدد زمنية مختلفة (قديمة، معاصرة،...).
- أن تكون مشكولة بالكامل.
- أن تكون مُعلَّمة tagged، أي توضع لكل كلمة معلومات مثل النمط اللغوي للكلمة POS وبنيتها الصرفية.

2- معايير تقويم المشكّل الآلي للنصوص:

بالرجوع إلى أنظمة التشكيل الآلي التي عرضناها آنفاً، وإلى نتائجها، يُمكننا أن نستنتج مجموعة من معايير التقويم التي تتيح لنا المقارنة بين نظم التشكيل الآلي. تتعلق بعض هذه المعايير بالوظائف نفسها ومدى تحقيقها للهدف المطلوب منها، في حين يعدُّ بعضها الآخر من عوامل الجودة التي قد تتعلق بالبيئة أو بالنظام نفسه (غير وظيفية). نورد فيما يأتي هذه المعايير وفق التصنيف المقترح:

أ- المعايير المتعلقة بالوظائف:

- الدقة وتقاس عموماً على مستويين، الأول يأخذ تشكيل أحرف الكلمة جميعها بالحسبان، والثاني يستثني تشكيل الحرف الأخير. وفي كل من

هذين المستويين، تقاس الدقة كما أسلفنا بالمؤشرين الآتيين:

- o نسبة الخطأ بالكلمات مع الحرف الأخير Word Error Rate (WER). وتُعبّر عن نسبة الكلمات المشكّلة خطأً إلى مجموع كلمات النص.
- o نسبة الخطأ بالتشكيل Diacritics Error Rate (DER). وتعبّر عن نسبة عدد الأحرف المشكّلة خطأً إلى مجموع أحرف النص.

- تقيس بعض المنهجيات المعتمدة على مدونات التدريب المؤشرين السابقين ولكن بالنسبة إلى الكلمات التي لم يجرِ التدريب عليها (أي ليست في مدونة التدريب):
  - o نسبة الخطأ بالكلمات غير الموجودة في مدونة التدريب Unseen Word Error Rate (UWER).
  - o نسبة الخطأ بالتشكيل للكلمات غير الموجودة في مدونة التدريب Unseen Diacritics Error Rate (UDER).

ب- المعايير المتعلقة بجودة الوظائف وبالبيئة:

- السرعة.
- بيئات العمل (على حاسوب مستقل، على الوب،...).
- استقلالية النظام عن نظم التشغيل.
- استقلالية النظام لتشغيله مع تطبيقات أخرى.
- إتاحة خيار التشكيل مع أواخر الكلمة أو من دونها.
- قابلية تحسين الأداء مستقبلاً.
- اعتماد المكونات في تصميم النظام.
- قابلية استدعاء خدمات النظام من تطبيقات أخرى.
- إتاحة المصدر أم لا.
- قابلية تخصيص النظام لنصوص في مجالات محددة.
- استقلال البرنامج عن الموارد اللغوية.
- حجم التخزين.
- سهولة الاستخدام.

V. منهجية العمل:

نظراً إلى وجود العديد من التجارب في مجال التشكيل الآلي، وبهدف توظيف هذه الخبرات في مشروع عربي يطمح إلى بناء مشكل آلي عربي مفتوح المصدر، جرت الدعوة إلى المشاركة في فريق عمل[1] يضم الباحثين الراغبين في مشاركة خبراتهم في التشكيل الآلي.

نعرض فيما يأتي مقترح خطة أولية، يسير وفقها فريق العمل المذكور آنفاً، لإنشاء مشكل آلي مفتوح المصدر، تأخذ بالحسبان التجارب البحثية السابقة، وتتضمن المراحل الآتية:

1- إنشاء مدونة مشكولة تشكيلاً كاملاً: بحيث يمكن استخدامها في الطرائق الإحصائية، وكذلك لإجراء عمليات التقويم بالاعتماد عليها (يمكن تجميع ما هو متوافر من مدونات مفتوحة، كما يمكن توسيع هذه المدونة بإضافة نصوص أخرى يجري تشكيلها وتدقيقها من قبل خبراء لغويين). وقد جرى التنسيق مع د. المعتز بالله السعيد على استخدام المدونة التي قام بتجميعها وهي مشكولة تشكيلاً كلياً وتتألف من نحو 13.5 مليون كلمة وتحتوي على مؤلفات قديمة ومعاصرة.

2- إنشاء نظام آلي لتقويم المشكلات الآلية: يأخذ هذا النظام بالحسبان مجموعة المعايير الكمية القابلة

[1] قامت ألكسو (المنظمة العربية للتربية والثقافة والعلوم) بالمبادرة إلى طرح مشروع عربي لبناء مشكل آلي مفتوح المصدر.

للقياس من المعايير المذكورة في الفقرة السابقة، كما يأخذ بالحسبان أيضاً أساليب التشكيل العديدة التي تعتمدها نظم التشكيل المختلفة (كأن يعتمد إحدى المشكلات عدم وضع حركة السكون فوق الأحرف، أو عدم تشكيل أحرف المد، ...)

3- إجراء تقويم للمشكلات الآلية التي يسمح مالكوها بفتح مصدرها (مشكّل مدينة الملك عبد العزيز [5]، مشكّل جامعة محمد الأول [13]، مشكّل Safadi [9]، وغيرها).

4- إجراء تعديلات على المشكلات الآلية المختلفة وفقاً لنتائج التقويم بهدف تحسين هذه النتائج، ويمكن إنشاء مشكل آلي يقاطع نتائج أكثر من مشكل.

5- تقويم النظم المختلفة في نهاية المشروع بهدف تحديد النظام ذي التقويم الأعلى الذي سيجري اعتماده لفتح مصدره.

6- إعداد توثيق موسع لهذا النظام بهدف جعله مفتوح المصدر وقابلاً للاستثمار بسهولة.

IV. الخاتمة:

عرضنا في هذه الدراسة منهجيات التشكيل الآلي للنصوص العربية التي استقصيناها من البحوث الحديثة المنشورة منذ بداية العقد الماضي وحتى الآن، ووجدنا أنها تندرج في صنفين اثنين: بحوث تعتمد المعلومات اللغوية التي توفرها المحللات الصرفية والنحوية والأنظمة التي تعالج الدلالة، وبحوث تعتمد المعلومات الإحصائية التي تستخرج من المدونات. ولعل التأمل العميق في بحوث الصنف الأول يبيّن أنه لابدَّ من استخدام الإحصائيات المستخرجة من المدونات في بعض مراحل المعالجة.

بناء على ذلك يمكن القول إنَّه لا يكفي الاهتمام ببناء نظم تشكيل آلي جيدة، بل لا بدَّ من بذل الجهود لبناء مدونات كبيرة الحجم ومشكولة ومعلّمة تضم نصوصاً في موضوعات مختلفة، وإتاحة هذه المدونات كمدونات معيارية لتقويم نظم التشكيل الآلي.

جرى الاعتماد على البحوث المجراة في مجال التشكيل الآلي لوضع خطة عمل أولية لمشروع يهدف إلى بناء مشكل آلي مفتوح المصدر تحت رعاية ألكسو، إذْ تشارك فيه جهات بحثية مختلفة من سورية، ومصر، والسعودية، وتونس، والولايات المتحدة، والمغرب.

VII. مسرد المصطلحات

محوّلات منتهية الحالات مثقلة	weighted finite state transducers
نموذج احتمالي مولِّد	generative probabilistic model
المعتمدة على السمات اللغوية	Language Feature Centered
المعتمدة على المعطيات	Data-Centered

المراجع*

[1] N. Habash, O. Rambow, 2007, "Arabic Diacritization through Full Morphological Tagging", Proceedings of 8th Meeting of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies Conference.

[2] M. Rashwan, M. Al-Badrashiny, M. Attia and S. M. Abdou, 2009, "A Hybrid System for Automatic Arabic Diacritization", Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April 2009.

[3] M. Maamouri, A. Bies, and T. Buckwalter. 2004. The Penn Arabic Treebank: Building a large-scale annotated arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

[4] R. Nelken and S. M. Shieber. 2005. Arabic Diacritization Using Weighted Finite-State Transducers. In Proc. of the ACL 2005 Workshop On Computational Approaches To Semitic Languages, Ann Arbor, Michigan, USA.

[5] M. Elshafei, H. Almuhtasib and M. Alghamdi, 2006, "Machine Generation of Arabic Diacritical Marks", The 2006 World Congress in Computer Science Computer Engineering, & Applied Computing . Las Vegas, USA.

[6] J. Giménez and L. Márquez, 2004, SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

[7] I. Zitouni, J. S. Sorensen, and R. Sarikaya. 2006. Maximum Entropy Based Restoration of Arabic Diacritics. In Proc. of the 4th Annual Meeting of ACL, Australia.

[8] M. Attia, M. Rashwan, 2004, A Large-Scale Arabic PoS Tagger Based on a Compact Arabic PoS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, Proc. of the Arabic Language Technologies & Resources Int'l Conf.; NEMLAR, Cairo.

[9] H. Safadi., O. Al Dakkak and N. Ghneim, 2006, "Computational Methods to Vocalize Arabic Texts" 2nd Workshop W3C, Herakilon, Greece.

[10] A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Conf. on Empirical Methods in NLP.

[11] O. Emam and V. Fischer. 2005. Hierarchical Approach for the Statistical Vowelization of Arabic Text. Technical report, IBM Corporation Intellectual Property Law, Austin, TX, US.

[12] T. Schlippe, T. Nguyen, and S. Vogel, 2008, "Diacritization as a Machine Translation Problem and as a sequence Labeling Problem", The 8th Conference of the Association for Machine Translation in the Americas, Waikiki, Hawaii, 270-278.

[13] M.S. Ryan, and G.R. Nudd, 1993, The Viterbi algorithm, Technical Report. Department of Computer Science, Coventry, UK.

[14] A. Mazroui, A. Mezian, A. Lkhwaja, M. weld Bebah, A. Bodlal, 2010, "A Statistical Approach for Arabic Diacritization", Enriching Arabic Digital Content Workshop, in Arabic, Damascus-Syria.

* تاريخ ورود البحث إلى مجلة جامعة دمشق 2012/6/14