هذه الطبعة
إهداء من المركز
ولا يسمح بنشرها ورقياً
أو تداولها تجاريا

اللسانيات العربية

The Arabic Linguistics Journal

مجلة علمية محكمة تصدر عن مركز الملك
عبدالله بن عبد العزيز الدولي لخدمة اللغة العربية
العدد ١٢ ربيع الآخر ١٤٤٢هـ - يناير ٢٠٢١ م

◆ أين شومسكي اليوم من مسألة تطور اللغة؟

◆ البنية الإبلاغية ووصف العربية: المحدَّث عنه والمحدَّث به. نموذجا

◆ العلامة الإعرابية في الفعل المضارع في ضوء نظريّة معنى-نصّ: الفعل المضارع الواقع بعد (حتّى) أنموذجًا

◆ معنى الجمع في المركّب بالواو

◆ الاستعارة المقترضة - المعجم السياسيّ أنموذجا مقاربة معجميّة عرفانيّة

◆ (حَسْبُكَ) - دراسة تركيبيّة في ضوء المَقُولات والقرائن

◆ الحقول والمجموعات الدلالية في نصوص مجال الاقتصاد: دراسة على مدونة حاسوبية

◆ تعليم القواعد النحوية بالاعتماد على أمثلة المدونات اللغوية (المفعول لأجله نموذجًا)

◆ قراءة في كتاب الدكتورة أفراح عبدالعزيز التميمي: نظام آلي للتقطيع والتوسيم النحوي العربي

◆ عبدالعزيز العصيلي العالم الذي افتقدته اللغويات التطبيقية

معامل التأثير والاستشهادات المرجعية للمجلات العلمية العربية أرسيف ARCIF رقم (٠,٢٣٥٣).

قراءة في كتاب الدكتورة أفراح عبدالعزيز التميمي: نظام آلي للتقطيع والتوسيم النحوي العربي

د. سلطان المجيول (*)

صدر هذا الكتاب في طبعته الأولى عام ٢٠٢٠م عن دار كنوز المعرفة للنشر والتوزيع (عمّان)، وتضمّن بعد المقدمة خمسةَ فصول: التوسيم النحوي، ومكونات بناء نظام التوسيم النحوي الآلي، ونظام التوسيم النحوي الآلي المقترح وتطبيقاته، وتطبيق المنهج (التوسيم النحوي بفروعه الثلاثة: الوسومات الرئيسة، والوسومات الفرعية، والوسومات الموسّعة بالخصائص التصريفية)، ثم النتائج والتوصيات.

أولًا: قراءة أولى

كان نظامُ ستانفورد للتقطيع والتوسيم النحوي حاضرًا في أثناء قراءتي لفصول هذا الكتاب. وعلى الرغم من تعدد أنظمة التوسيم النحوي العربي، كوسوم بيز Bies وبادت PADT ومدى MADA وأميرا AMIRA ومداميرا MADAMIRA، فقد كان عاملُ التناسخ reproducible (إمكانية تطبيق نظام الوسوم تطبيقًا مفتوحَ المصدر على البيانات الخاصة به) وعاملُ التناتج replicability (إمكانية تطبيق نظام الوسوم تطبيقًا مفتوحَ المصدر على بيانات أخرى) اللذان تتميز بهما وسوم ستانفورد متاحيْن في نظام التقطيع والتوسيم النحوي الذي عملت الدكتورة أفراح التميمي على تأطيره وتصنيفه وتطبيقه واختبار أدائه. وتتوفر أنظمة الوسوم بالشكل البحثي

* أستاذ اللغويات الحاسوبية والتطبيقية بجامعة الملك سعود.
أرسل البحث بتاريخ ١٧/ ١١/ ٢٠٢٠م، وقبل للنشر بتاريخ ١٦/ ١٢/ ٢٠٢٠م.

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

في عدة أنظمة، لكنني لا أجد توفر تناسخها أو تنائجها إلا في نظام ستانفورد أو في النظام الذي تقترحه مؤلفة هذا الكتاب. وكان عاملُ التناسخ والتناتج متاحًا في الكتاب، وحاضرًا في الموقع المصاحب له في github، كما وفّرت المؤلفة أيضًا جزءًا من البيانات لمعاينة أداء النظام على المدونة الفرعية (الإصدار الرسمية): https:// .github.com/AfrahAltamimi/TamimiTags

ثانيًا: قراءة ثانية

استعرضت المؤلفة في الفصل الأول الأسس العلمية والمنهجية للتوسيم النحوي، وأنواعه، وأهميته. واستفاضت في تعداد أهمية التقطيع والتوسيم، وهي أهمية لا تقتصر على تعلّم الآلة للحالات النحوية الإعرابية للكلمات العربية فحسب، بل تشمل فوائد هذا التعلّم الموجّه supervised الذي يُعوّل عليه في فهم المرحلة القبلية (الخصائص التصريفية–التركيبية) وفهم المرحلة البعدية (السمات الدلالية). وما دامت الأنحاء اللغوية المشهورة (التوليدية، والبنيوية، والوظيفية) ذات نزعة تركيبية، فلا عجب أن يكون منطلق فهم اللغة كامنًا في طبيعة التجاذب attraction والتنافر repulsion بين أقسام (أحياز) الكلام. وتتطلب هذه الطبيعة من العقل اللغوي البشري تحليلا معمقًا في معظم الحالات، فكيف بحال العقل الآلي الذي يحتاج إلى مزيد من التعلّم النوعي لأقسام الكلام في اللغة العربية.

قدّمت المؤلفة في الفصل الثاني مكونات بناء أي نظام للتوسيم النحوي الآلي، مع الوقوف على شواهد لأنظمة موجودة، وحدّدت أغراض التوسيم المبني على إطار عينة من عشرة آلاف كلمة من المدونة العربية، كما نبّهت القارئ إلى أهمية تنويع إطار المدونة عند تدريب أنظمة التوسيم النحوي؛ لأن هذا التنويع يكشف بعض المتشابهات التي يصعب حل مشكلتها، كالوقوف على كلمة (دَنَا) و(دُنا)، واختلاف الثانية عن الأولى بـ (أل) التعريف. هنا تكمن أهمية التجريد tokenization وبناء نظامٍ للتقطيع من أجل أن تتعلّم الآلة كيفية التمييز بين هذا النوع من المتشابهات، ولكي تتقرّر آليا فعليةُ الكلمة الأولى، وتتقرّر آليا اسميةُ الكلمة الثانية. ويتطلّب هذا الإجراء بيانات دقيقة التنويع، ويتطلب تدريبًا لغويًا، ويتطلب اختبارًا لبيانات

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

خفيّة لمعرفة أداء تعلّم الآلة. وسيجد القارئ أن المؤلفة قد تصدّت في الفصل الأول والفصل الثاني لمعظم الإشكالات التطبيقية التي تواجه أنظمة التوسيم النحوي، ثم فصّلت في الفصل الثالث والفصل الرابع النظام الآلي الذي اقترحتْه للتقطيع والتوسيم النحوي.

ثالثا: قراءة ثالثة (التوسيم النحوي بين العقل اللغوي والعقل الآلي)

بلورت المؤلفة إطارا لتوسيمٍ نحوي يفتح رؤية لغوية جادّة. وأريدُ قبل البدء في طرح القراءة الثالثة أن أنوه إلى احتماليْن أساسيّين من منظور العقل الحاسوبي الرياضي المحض. الأول: الآلة لا تجابه العقل البشري النوعي المضطرب في التفسير؛ لأن العقل الحاسوبي التقني ذو إطار كمّي دقيق غير متناهٍ. فأيُّ عقل لغوي عربي ينتظر من الآلة أن تكون عصا سحرية من الناحية النوعية، وهو بذلك يكون بعيدًا كلّ البعد عن طبيعة الدوائر التطبيقية في تعلّم الآلة، فتعلّم الآلة أو التعلم العميق (الفرع منه) للغة العربية ليس مواقف تفسيرية، وإن كانت بعض المشاريع فيه محسوبة على هذا النوع من المواقف، كالتي ترى أن تعلّم الآلة بالتلقين هو الذكاء الاصطناعي. وما دام الذكاء البشري ليس معتمدًا على التلقين، لكون التلقين جمودًا، فإن الذكاء الاصطناعي ليس تلقينا ولا يكون بجعل الآلة محاكاةً. ولو أن فريقًا ابتكر لنا روبوتا يُنشدُ لنا قصيدةً عربية، ثم نجعل منه ذكاءً اصطناعيا، فهذا ليس من الذكاء الاصطناعي بشيء، إنما هو مجرد محاكاة. قد يخلط بعض اللغويين وبعض الحاسوبيين بين محاكاة الآلة للعقل البشري وبين تعلم الآلة والتعلم العميق، كما قد يخلط البعض بين إفادة الآلة بطبيعة التحليل اللغوي (تعلّم الآلة) والاستفادة من الآلة في التحليل اللغوي (التطبيقات).

ولو جئنا إلى معالجة اللغة العربية، فالمشروع العالمي المتعلق بالتقطيع فالتوسيم هو مشروع ستانفورد؛ لأنه مشروع مؤسسي دعمًا ونشرًا. وقد استعملتُ هذا النظام أكثر من ثمان سنوات، وإني أراه مع ذلك مميزا. لكن، ماذا عن النظام الآلي للتقطيع والتوسيم النحوي العربي المطروح في الكتاب؟ هناك أطرٌ يحتاج إلى أن يتعمّق فيها العقل اللغوي قبل أن يقيّم أداء الذكاء الاصطناعي في معالجة اللغة العربية، فالعقل

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الآلي يُدركها بامتياز كمّيّ هائل، ويتطلّب من العقل اللغوي أن يتوقّف عند النواتج النوعية لبعض مخرجات هذا الإدراك الآلي، فيُراجع الأخطاء المعدودة ليصححها، أو يقلصّ من التصنيف النحوي، أو يزيد في التصنيف النحوي، شرطَ أن يُحقّقَ هدف تحسين أداء النظام الآلي بشكل أعلى.

رابعًا: قراءة رابعة (إمكانات النظام المقترح)

سأذكر في الآتي جهد المؤلفة في النظام المقترح الذي قامت ببنائه لتقطيع النص العربي وتوسميه نحويًّا. حدّدت المؤلفة نظام التوسيم النحوي الآلي المقترح وتطبيقاته في الفصل الثالث، فجرّدت الكلمات العربية في إطار عينة مدونتها المختارة، المكونة من عشرة آلاف كلمة، وإجراء التجريد tokenization إجراءٌ أساس، وقد أخذت المؤلفة في الحسبان عدم إجراء بعض التجريد الذي من شأنه إحداث إخلال في توسيم الكلمة المجرّدة، فلم تفصل التاء في (طلحة) وفي (حقيبة)، ولم يطَلْ إجراء التقطيع الضمائرَ المتصلة للمتكلم، والمخاطب، والغائب، والمفرد، والتثنية، والجمع، والتذكير، والتأنيث، كما لم يطَلْ حروف المضارعة (أنيت)، وذلك لسببين؛ أولهما: مجانبة ما سيطرأ من خطأ عند التوسيم النحوي الآلي، وثانيهما: تسهيل وضع نظامٍ للتوسيم النحوي الموسّع بالخصائص التصريفية. وبذلك، حدّدت المؤلفة متغيرات التقطيع القائمة على المبنى التقسيمي النحوي، وذلك في همزة الاستفهام والتسوية، وحرف الجر أو القسم (الباء)، وتاء القسم، وحرف الاستقبال (السين)، و(عن) المدغمة نونها في ميم (من) و(ما)، و(من) المدغمة نونها كما في (ممن، ومما، ومِمّ)، و(في) المتصلة بـ (من)، وفاء الشرطية، وفاء السببية والاستئنافية والعاطفة، وكاف الجر، ولام (اللام) الجر، ولام الأمر، ولام التأكيد، وهاء التنبيه، وحروف الجر (الباء، وفي، واللام، وعلى، وإلى، وحتى) المتصلة بـ (ما الاستفهامية المحذوفة الألف)، وإذ الظرفية المتصلة بـ (حين)، و(عند)، و(وقت).

رَسَمتْ المؤلفة إطار التجريد الذي به رسّمت متغيرات التقطيع؛ فبعد تأطير متغيراتِ التقطيع، وضعت إطار الوسوم النحوية الرئيسة والفرعية والموسّعة بالخصائص التصريفية. ومن شأن هذه الإجراءات الثلاثة المتتابعة أن جعلت

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

التحليل الآلي للنص العربي المتمثل في إطار عينة المؤلفة واضحًا، وهذا ما لا يُمكن لنا أن نراه في وسوم أخرى غير وسوم المؤلفة. وجعلت المؤلفة الوسوم الرئيسة متمثلة من أقسام الكلام التي وضعها تمام حسان في كتابه (اللغة العربية معناها ومبناها). ومن منطلق هذه الأقسام، وضعت المؤلفة الوسوم النحوية الفرعية وفقَ تفاصيل التقسيمات الفرعية لأقسام الكلام السبعة. أما في الوسوم الموسّعة بالخصائص التصريفية، فقد بلغ ما صنّفته المؤلفة ٣٩٨ وسمًا، وهي عبارة عن وسوم رئيسة وفرعية مركبّة يدويا بشكل يصف الحالات التصريفية-التركيبية التي جاءت عليها الكلمات في إطار العينة.

في الفصل الرابع، يُحسب للمؤلفة وضع تقييمين اثنين؛ أولهما: تقييم لغوي، وثانيهما: تقييم آلي. وطبّقت المؤلفة نموذج الحقول العشوائية المشروطة conditional random fields، وهو من أذكى النماذج الآلية في تقييم أداء التصنيف آليا. وللمؤلفة دور جديد على المهتمين بالتوسيم النحوي، نجده في قياس الانتقالات الاحتمالية (صفحة ٢٦٩ و٢٧٠). وتتعلق هذه القياسات بأداء النظام الذي اقترحته المؤلفة على مستوى الوسوم الرئيسة والفرعية والموسعة. ومن النادر- في تعلّم الآلة ببيانات ذات خصائص متشعبة ومتباينة كالخصائص التصريفية الموسّعة التي بلغت ٣٨٩ وسمًا- أن يتحلى هذا التعلّم بالعشوائية المنضبطة؛ لكن نسب أداء النظام ظهرت بشكل شبه متطابق في أثناء التقييم في كتاب المؤلفة. وسأقف هنا عند نسب الصحة accuracy والدقة precision والاسترجاع recall، فيما يُعرف بمصفوفة الارتباك confusion matrix.

أشارت المؤلفة إلى أن معظم أنظمة التوسيم النحوي المشهورة،كنظام ستانفورد ومدى وأميرا، لا تُفصح عن قيمة الدقة والاسترجاع. ومما يُحسب للمؤلفة في هذا السياق عرضها نِسب الصحة والدقة والاسترجاع لأنظمة الوسوم النحوية المقترحة. فقياس الصحة يعني النسبة المئوية للوسوم النحوية الصحيحة عند تصنيف الكلمات العربية في البيانات، والاكتفاء بالصحة accuracy لا يكفي لإيضاح جودة النظام وأدائه عندما تكون أصناف أقسام الكلام متباينة وغير متزنة. ولو أن أقسام الكلام

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

تتوازن في النص العربي من حيث الورود توازنًا عدديًا، كأنْ تحتل الوسوم الاسمية الربع الأول من النص، وتحتل الوسوم الفعلية الربع الثاني، وتحتل وسوم الصفات الربع الثالث، وتحتل وسوم البقية الربع الرابع، لكانت نسبة الصحة كافية ومقبولة نوعا ما. لكن النص اللغوي لا يكون كذلك بطبيعة الحال، وعليه تجيء ضرورة الإبلاغ عن نسبة الدقة والاسترجاع، لكون الدقة كاشفًا عددَ التصنيفات الإيجابية الخاطئة، وكون الاسترجاع كاشفًا عددَ التصنيفات السلبية الصحيحة، ثم يزيد قياس F1 تقييم كل من الدقة والاسترجاع، ليُعطي بذلك القيمة النهائية لأداء نظام التوسيم بشكل كامل.

كلٌ من نسبة الصحة والدقة والاسترجاع وF1 شبه متطابق في المجموعات التوسيمية النحوية الثلاث التي اقترحتها المؤلفة، فنِسبُ الوسوم الرئيسة فوق التسعين، ونِسبُ الوسوم الفرعية فوق الثمانين، ونِسبُ الوسوم الموسّعة فوق السبعين. وعلى الرغم من أن الحاسوبيين يرون أهمية التقييم التقني من حيث السرعة والأداء وصحة النموذج بعد اختباره، فإن التقييم اللغوي يضع الحاسوبيين في بعض الأحيان عند مفترقٍ من الطرق. وتكمن أهمية التقييم اللغوي في فحص البيانات اللغوية التي دُرّب بها النظام، ويَنْدرُ على حد علمي أن نجد أبحاثًا حاسوبية تُتيح البيانات اللغوية، فلا نجد إلا نسبَ الصحة، وأمثلةً مختارة في جداول منشورة في أوراق علمية، مما يتطلّب أن تكون هذه البيانات متاحة لغرض التقييم اللغوي الذي بدوره سيُحسّن كثيرا من الأداء الآلي (التقني).

خامسًا: قراءة خامسة (ختامية)

تبدأ المرحلة القبلية في معالجة اللغة العربية بالتقطيع والتوسيم. ومعظم المعالجَات -إن لم يكن كلها- تتطلب تطبيق التقطيع والتوسيم من أجل فهمٍ أعلى لطبيعة بنيات النص العربي التركيبية والدلالية. ومن المعالجات التي تحتاج إلى معالجة قبلية كالتنقيح والتقطيع والتوسيم، استخراج الحقائق من المعلومات في النص، وتحليل المشاعر، والترجمة الآلية، ونمذجة الموضوع، إلخ.

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

يلجأ معظم المطوّرين في اللغويات الحاسوبية إلى نظام ستانفورد للتقطيع والتوسيم، لكونه يُقدّم معالجة قبلية مقبولة. ومن زاوية واقع التطوير في المعالجة، يبقى هذا النظام مشهورًا ليس لأنه نظام يتحلّى بسمات تحليلية عميقة، بل لأنه نظام مؤسسي دولي وقويّ في معالجة اللغة الطبيعية. ومن سمات الابتكار أن يكون العمل قابلا للتطوير، وكل عمل يسعى إلى بسط آفاق البقاء والتطور مستقبلا يحتاج إلى أن يُتيح فيما انتهى إليه أبوابًا مفتوحة (غير مغلقة) لعدد من الأفكار الابتكارية. ومن شروط الابتكار أن تكون أبواب الأفكار مشرّعة على الدوام، وأن تكون الأسئلة والفرضيات ذات أرجحيات احتماليّة تتوافق مع تفوّق حجم التدريب والاختبار والنتائج. وأظن أن الإطار الذي اقترحته المؤلفة سيعيد النظر في اتجاهات التوسيم النحوي للغة العربية، كما سيعيد النظر في معالجات النص العربي القديم، والنص العربي الحديث بمستوييه النموذجي (الأدبي والأكاديمي والصُّحُفيّ) واللّهجي.

سيُعيد النظامُ المقترح على صعيد مجموعة الوسوم الرئيسة النظرَ في ضرورة الخروج من قيد التوسيم الآلي البسيط القائم على ثلاثية التقسيم المنطقي (اسم وفعل وحرف).

وفي نظام الوسوم الفرعية النحوية المقترحة من المؤلفة مزية لا تقتصر على أدائه الآني، بل تظهر في تطبيقاته اللاحقة على بيانات لغوية متنوعة بتنوع الزمن، لأغراضٍ تتحدّد بتحسين الأداء. وأفترضُ هنا أن هذه الوسوم الفرعية ستتحسن لو طُبّقت على بيانات كبيرة متنوعة بتنوع الزمن، لتظهر – على سبيل الافتراض– أربع مجموعات توسيمية لهذه الوسوم الفرعية: مجموعة وسوم النص القرآني، ومجموعة وسوم النص العربي النثري القديم، ومجموعة وسوم النص العربي الشعري القديم، ومجموعة وسوم النص العربي المعاصر.

أما الوسوم الموسعة بالخصائص التصريفية، فتكون لإجراءات تحسين الأداء على مجموعة بيانات متنوعة بالتنوّع النموذجي واللهجي، وبناء على هذه القراءة، سأفترض أنها ستكون فاتحةَ بابٍ لمعالجة مشكلات المزجيّة fusion التي تتصف بها اللغة العربية بامتياز في تراكيبها النموذجية واللهجية المعاصرة.

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً

من الاتجاهات المستقبلة في التوسيم النحوي: التعلم العميق. ويتطلب هذا النوع من تعلّم الآلة بيانات تدريب ضخمة، فتعلّم الآلة بنماذج الشبكات العصبية على أقسام الكلام، وتدريبها على ذلك، في بيانات تصل إلى مئات الآلاف، يُعدّ أساس أي مشروع في هذا المجال. ولو حقّقت المؤلفة هذا على مستوى الوسوم الرئيسة ثم على مستوى الوسوم الفرعية، فسيكون مشروعًا عميقًا في مجال معالجة اللغة العربية الطبيعية.

يبَقى على المؤلفة بعد هذا المنجز العميق في مشروعها أن تسعى إلى تصميم واجهات استخدام لكل نوع من هذه الوسوم، وإظهار تطبيقات قائمة بذاتها، لتتيح للغويين وغير اللغويين ممن لا يتعاملون مع لغات البرمجة الاستفادة من نظام التقطيع والتوسيم النحوي المقترح في أبحاثهم ومشاريعهم ذات العلاقة.

هذه الطبعة إهداء من المركز ولا يسمح بنشرها ورقياً أو تداولها تجارياً