International Journal of Arabic Linguistics

نحو تحسين أداء نموذج التميمي للوسم النحوي الآلي للغة العربية

أفراح عبد العزيز حمد التميمي
جامعة الإمام محمد بن سعود الإسلامية
المملكة العربية السعودية
aahaltamimi@imamu.edu.sa

ملخص

يعد التوسيم بأقسام الكلام من أهم المهام في معالجة اللغات الطبيعية. وثمة العديد من الأعمال والمشاريع والجهود المبذولة في هذا الميدان، غير أنها لا ترقى للمستوى المطلوب بسبب العديد من الإشكاليات سواء على مستوى التنفيذ أو على مستوى الأداء. وتهدف هذه الورقة إلى تحسين أداء النموذج الأولي للتوسيم النحوي بمجموعة وسوم التميمي الأساسية، المبني بالاعتماد على مدونة موسمة يدويا باثني عشر وسما نحويا (الاسم – الصفة - الفعل – الضمير – الظرف -الخالفة – الأداة- علامة الترقيم – اختصار- كلمة أجنبية -رمز- رقم)، وذلك بإعادة تدريب النموذج model retraining وإضافة 46,811 كلمة فعلية على امتداد أزمنة العربية وأمكنتها وموضوعاتها. وفيما قد حقق النموذج الأولي درجة صحة 91,58%، حقق النموذج المحسن 93,97%.

Abstract

One of the most important tasks in natural language processing is POS tagging. There are many works, projects and efforts made in this field, but they are not up to the required linguistic level due to many problems that arise at the level of implementation or performance. This paper aims to improve the performance of the initial CRF model of the POS tagging with the basic Tamimi tags based on a hand-tagged corpus with twelve grammatical tags (noun - adjective - verb - pronoun - adverb - interjection - practical - punctuation - abbreviation - foreign word - symbol - number), by the model retraining method and adding 46,811 tokens throughout the Arabic eras, its regions, and domains. The performance has developed as the initial model achieved an accuracy score of 91.58%, and the enhanced model achieved a score of 93.97%.

الكلمات الدلالية: التوسيم النحوي - النموذج الأولي - مجموعة وسوم - مدونة التدريب - مدونة الاختبار - إعادة تدريب النموذج - خوارزمية

Keywords:

POS tagging - initial model - tagset - train corpus - test corpus - retraining model - algorithm

© المجلة الدولية للسانيات العربية 7(1)، ص. 61-79

1 المقدمة

يقصد بالتوسيم النحوي إسناد الرموز التي تشير إلى نوع الكلمة (فعل، اسم... إلخ) لكل وحدة معجمية في النص (McEnery, et al., 2006, p. 34). ويسمى أيضا التوسيم النحوي 'grammatical'، أو التوسيم التركيبي الصرفي 'morpho-syntactic'، أو التوسيم بأقسام الكلام 'POS Tagging'. وهو يختلف عن التوسيم بالتحليل التركيبي 'parsing' الذي يستعمل التشجير والأقواس (McEnery & Wilson, 2011, p. 46).

وثمة العديد من خوارزميات التصنيف 'classification' المستعملة في تعلم الآلة 'machine learning'؛ لبناء أنظمة التوسيم النحوي الآلية 'taggers'. وهي وإن اتفقت معا في عملها، إلا أنها تختلف في نظامها الرياضي والتقني عند حل المشكلات التصنيفية. ومن تلك الخوارزميات: خوارزمية الحقول العشوائية المشروطة 'Conditional Random Fields (CRF)'. وهي خوارزمية من خوارزميات تعلم الآلة الموجه 'supervised' المستعملة لبناء نماذج تمييزية 'discriminative models' احتمالية، تتنبأ بوسوم كلمات متتالية، مع أخذ السياق بعين الاعتبار (D. Lafferty, et al., 2001, p. 282). والنموذج 'model' في مجال تعلم الآلة، هو نتاج الخورازمية بعد تدريبها على البيانات، ويمثل ما تعلمته الخوارزمية من البيانات، وما تم حفظه بعد عمل الخوارزمية على بيانات التدريب. وهو عبارة عن القواعد والأرقام وأي هياكل بيانات أخرى خاصة بالخوارزمية، ومُتطلَّبة لإجراء عمليات التنبؤ. وهكذا، ينتج عن خوارزمية التصنيف CRF' نموذجا يمكن أن يتنبأ بصنف 'class' واحد (أو تصنيفات متعددة) لكل عنصر في البيانات، ويمكن حفظه واستعماله لاحقا.

وتتطلب نماذج تعلم الآلة العالية الأداء بيانات عالية الجودة. وفي مجال التوسيم النحوي العربي، يظهر في نماذج التوسيم النحوي الآلية مشكلات مختلفة ذات صلة بالبيانات 'corpus' ومجموعات الوسوم 'tag set'. فتلك النماذج لم تدرب على نصوص متنوعة من الفصحى على امتداد عصورها وأوعيتها. ومنها نماذج لا تراعي قواعد العربية لا في مجموعات وسومها ولا فيما تضمنه هذه الوسوم. ووجود نموذج للتوسيم النحوي ينطلق بوسومه من العربية، ولا يستل من لغة أخرى يضبط عملية التوسيم. بالإضافة إلى أن توفر مدونة عربية بمحتوى ممتد على عصور العربية وأوعيتها، وموسمة نحويا، يفيد الباحثين والمهتمين لغويا وحاسوبيا في أعمال متعددة.

ومن أساليب تحسين النموذج الأولي 'initial' في تعلم الآلة ما يعرف بمنهج إعادة تدريب النموذج 'retraining model' باستعمال بيانات التدريب الأولية التي تحدث بمجموعة بيانات إضافية (Klabjan & Zhu, 2020). وفي الظروف التي يصعب فيها جمع بيانات ضخمة مُمثِّلة وموسمَّة يدويا، وتكون البيانات متغيرة بتغيرات طبيعية لا يمكن التحكم فيها أو تقييدها، تأتي الحاجة لاستعمال هذا المنهج. وليس ذلك لتوفير الجهد والوقت وحسب، بل ولتحسين الأداء أيضا. ويعني منهج إعادة التدريب 'model retraining' في سياق تعلم الآلة، القدرة على تحسين النموذج بطريقة سلسة تراعي المهمات، وتوزيع البيانات المختلفة، مع الاستمرار في إعادة استعمال النموذج للمعرفة والمهارات التي اكتسبها، والاحتفاظ بها. ففي إعادة تدريب النماذج يُبنى فوق ما تَعلَّمهُ النموذج مسبقا، وهذا يعني دعم النموذج الأولي ليتعلم مع كل بيانات جديدة.

2 المشاريع ذات العلاقة

تلتقي المحاولات المسجلة للباحثين في أنظمة الموسمات النحوية العربية منذ بداياتها وحتى الآن مع عملنا الحالي في بعض ما تقدمه. ولكنها تختلف معه في الأسس التي بنيت عليها، والمبادئ التي انطلقت منها. فقد تنشّأت معظم المحاولات السابقة من تصورات لأقسام الكلام في لغات أخرى، وبعضها اعتمد التقسيم الثلاثي التقليدي العام الذي لا يكفي في أبحاث معالجة اللغة، دون إحاطة بما لدينا من نظريات لغوية عربية حديثة، قدمت تصورات لأقسام الكلام في العربية، وهو ما استند عليه النموذج المبحوث.

وإذا ما نظرنا في الجهود السابقة لعملنا، فإننا نجدها في اتجاهين هما: المدونات الموسمة، والأنظمة التي تستعملها هذه المدونات. والمدونات الموسمة إما مدونات موسمة توسيما يدويا، أو مدونات موسمة توسيما آليا بأحد أنظمة التوسيم المتوفرة وعلى نفس نوعية البيانات التي دربت عليها. وأما أنظمة التوسيم فإنها لا تخلو من أن تكون أنظمة مطبقة فعليا على مدونات لغوية لم تدرب على نوعية نصوصها، كنظام ستانفورد ‘Stanford’ (Toutanova, et al., 2003)، ونظام مدى ‘MADA’ (Habash, et al., 2013)، ونظام أميرا ‘AMERA’ (Diab, et al., 2004)، وأنظمة متطورة منها، نحو: ‘MADAAMIRA’ (Pasha, et al. 2014)، وجميعها أنظمة تعالج النص العربي الخام وفقا لمعيار البنك الشجري العربي لجامعة بنسلفانيا ذي النصوص الصحفية. أو أنظمة مقترحة نظريا، ولم تطبق. ومنها: مقترح الثبيتي (2012) الذي أشار فيه إلى قواعد في العربية ليست على إطلاقها، كالاعتداد بالتاء المربوطة في خاصية التأنيث، واعتماد الكلمتين: (في – عن) حرفي جر دون النظر في التباس الأول بالاسم (فَيّ)، والثاني بالفعل (عنَّ) عند غياب التشكيل. وكذلك مقترح الحاج (2013) المبني على الخصائص اللغوية لمفردات النص القرآني الحاج.

ووفقا لذلك يمكن تقسيم الجهود السابقة المتعلقة بموضوع الورقة إلى: مدونات موسمة يدويا، ومدونات موسمة آليا، وأنظمة توسيم نحوية آلية اختبرت على غير نوع النصوص التي دربت عليها. وقد كانت أبرز الأعمال المتعلقة بالمدونات الموسمة يدويا عمل خوجة (2001) الذي أسفر عن أول موسم نحوي آلي للعربية، عرف باسم الموسم الآلي بأقسام الكلام ‘Automatic Arabic POS-Tagger’ (APT) (Khoja, et al., 2001). وقد انطلقت في مجموعة وسومه من التقسيم الثلاثي العربي، فجاءت وسومه من خمسة أقسام أساسية هي: الاسم والفعل والأدوات وعلامات الترقيم والفَضْلات ‘residuals’. ثم وسمت يدويا بتلك الوسوم النحوية أربع مدونات هي: مدونة لصحيفة الجزيرة بتاريخ 3-3-1999 مكونة من 59 ألف كلمة، ومدونة لصحيفة الأهرام بتاريخ 25-1-2000 مكونة من 3 آلاف كلمة، ومدونة لصحيفة البيان القطرية بتاريخ 25-1-2000 مكونة من 5800 كلمة، ومدونة من المشكاة المصرية في العلوم الاجتماعية في أبريل 1999 ومكونة من 17 ألف كلمة. استعملت هذه المدونات لبناء معجم بكلماتها، ولتدريب نموذجها الهجين ‘APT’ الذي حقق نتائج وصلت دقتها إلى 86%. وبالنظر مثال نصي قدمته خوجة وآخرون (2001, p. 348) للكشف عن أداء نظام التوسيم النحوي الآلي المقترح، كشفتُ عن أخطاء في توسيمها بلغت تقريبا 50% وهي أخطاء صرفية تارة، ونحوية تارة أخرى، وقد تكون صرفية نحوية من ناحية ثالثة. فضلا عن الأخطاء المتعلقة بمجموعة الوسوم نفسها، كإدراجها للظروف في قسم الحروف على خلاف ما يراه أصحاب التقسيم الثلاثي. وحيث إن منطلق الدراسة حاسوبي وليس لغويا، فوقوع الأخطاء اللغوية متوقع. وفضلا عن تميز نموذجنا بانطلاقته اللغوية المتخصصة وانتمائه لكل عصور العربية، فقد درب موسم خوجة على مدونة تقتصر على لغة الصحافة العربية ونظائرها.

ومن المدونات الموسمة يدويا، مدونة البنك الشجري العربي. وهي مدونة أعدتها جامعة بنسلفانيا، وقدم القائمون عليها ثلاثة إصدارات كاملة للبيانات الموسمة صرفيا وتركيبيا (Maamouri, et al., 2004, pp.102-109). الإصدار الأول 166 ألف كلمة من صحيفة فرانس برس 'France Press'، والإصدار الثاني 144 ألف كلمة من صحيفة الحياة، والإصدار الثالث 350 ألف كلمة من صحيفة النهار. وقد استعملوا محلل باكولتر الصرفي كنقطة بداية للتوسيم الصرفي والتوسيم بأقسام الكلام لهذه المدونات بإصداراتها الثلاثة، حيث يوفر المحلل حلولا بالتشكيل التام متضمنة السوابق واللواحق والجذوع للكلمة الواحدة وأقسام الكلام. ومن 2002 إلى 2004 ظهرت هذه الإصدارات وأسست كبنك شجري عربي مكون من نصف مليون كلمة محشاة وموسمة باستعمال البنك الشجري الإنجليزي كمساعد. ومع نتائج تطبيق كل إصدار تعاد معالجة النتائج في الإصدار الذي يليه من خلال نظام يلبي متطلبات البنك الشجري ويسد ثغراته. وقد كانت النتائج تتحسن مع كل إصدار. ويضم نموذج التوسيم الخاص بمدونة البنك الشجري أكثر من 2200 وسم منها 114 وسما أساسيا. وتقدم هذه الوسوم مجموعة من الخصائص الصرفية: كالحالة الإعرابية 'case'، والموقع الإعرابي'mood'، والتعريف 'definiteness'، والجنس 'gender'، والعدد 'number'، وزمن الفعل'tense'، والجهة 'aspect'، والفعل من حيث بنائه للمجهول أو المعلوم voice. وقد عرض الباحثون (Maamouri, et al., 2004) مثالا مشروحا لمنهجية البنك العربي الشجري بالتركيز على بناء خاص في كل من التحليل الصرفي والتوسيم والتحليل التركيبي متبوعا بتفصيل وشرح كامل لعملية التوسيم، أشاروا فيه إلى خطوات عملية التوسيم، وأن المدونة التي طبقت عليها هذه الوسوم هي نصوص عربية صحفية غير مشكولة. وهذه المحاولة تشبه المحاولة التي سبقتها في استهدافها مجالا معينا من المدونات، حيث طبقت التوسيم على نصوص صحفية في فترة زمنية محددة. كما أنها تخالف عملنا في تناولها لمعالجات صرفية وتركيبية لم أتطرق لها في هذه الورقة.

وأشير أيضا في هذا السياق إلى مشروع مدونة القرآن الكريم، المورد اللغوي الموسَّم بطبقات متعددة من التحشية متضمنة التقطيع الصرفي أولا ثم التوسيم بأقسام الكلام والتحليل التركيبي باستعمال النحو العلائقي. ولقد كان الدافع وراء هذا العمل هو إنتاج مورد يمكّن من الحصول على مزيد من التحليل للقرآن الكريم. وتصف ورقة عمل قدمها نزار حبش في المؤتمر الدولي السابع لموارد اللغة والتقييم في مالطا (Dukes & Habash, 2010, pp. 2530-2536) منهجا حديثا في التوسيم الصرفي لعربية القرآن الكريم التي تختلف عن المستويات الأخرى للغة، حيث تعد لغة القرآن تحديا فريدا من وجهة نظر حاسوبية لاختلافها في عناصرها ومستوياتها اللغوية عن اللغة العربية الفصحى. واختلفت مدونة القرآن الكريم عن باقي المدونات في تبنيها مجموعة الوسوم من كتب النحو العربية التراثية وطرحها المدونة للمهتمين بالتوسيم عبر موقع خاص لها على شبكة الإنترنت. وقد نوقش في هذه الورقة كيف أن التحدي الفريد لعملية التحشية للقرآن الكريم حُلّ باستعمال أسلوب متعدد المراحل، وهذه المراحل تتضمن التوسيم الصرفي الآلي باستعمال التحقق اليدوي المزدوج، والتوسيم التعاوني عبر الإنترنت حتى من غير المتخصصين. وقد قيمت هذه العملية للتحقق من ملاءمتها للمنهجية المختارة. ورغم وجود الأخطاء اللغوية في تحديد الوسوم للكلمات، إلا أن هذا المشروع يتفق مع عملنا في اعتماده على مدونة وتوسيمها بأقسام الكلام، ويختلف عنها في تخصيصه لمستوى واحد من مستويات العربية هو المستوى الذي يمثله النص القرآني، وفي هدفه الخاص بعمليات تحليل النص القرآني.

أما ما يتعلق بالمدونات الموسمة آليا، فمشروع البنك الشجري العربي لجامعة كولومبيا (كاتِب CATiB) هو قاعدة بيانات للتحليلات النحوية في الجمل العربية (Habash & Roth, 2009). وقد اشترك في المشروع خمسة موسمين وسموا هذه المدونة المكونة من مجموعة

نصوص صحفية، من وكالات أنباء متعددة، في الفترة الزمنية ما بين 2002-2007. كما يشتمل هذا البنك الشجري على بنك شجري إضافي محول آليا من البنك العربي الشجري. وبنك (كاتب) أصغر من البنك العربي الشجري حيث يضم ستة أقسام فقط هي: أسماء الأعلام، وأسماء غير الأعلام، والأفعال، والأفعال من حيث البناء للمعلوم أو المجهول، والأدوات، وعلامات الترقيم. واستعمل بنك كاتب الأداة 'MADA&TOKAN' (Habash & Rambow, 2005) للتقطيع وللتوسيم بأقسام الكلام الذي يقدم أكثر من تحليل محتمل للكلمة الواحدة. فقُطِّعت المدونة ثم وسمت وحللت تحليلا نحويا آليا. وقد حقق التقطيع صحة بلغت 99.1%، فيما حقق التوسيم النحوي دقة بلغت 99.7%. ثم روجعت نتائجها وصححت يدويا. وتتفق هذه الورقة مع الدراسة الحالية في تناولها لمدونة لغوية، وتوسيمها توسيما نحويا آليا ثم مراجعتها يدويا، وتختلف معها في وعاء النصوص وزمنها.

ومن المدونات الموسمة آليا أيضا مدونة 'arTenTen' (Yonatan, et al., 2013). وهي مدونة عربية من مجموعة مدونات 'TenTen'. وقد جمُعت نصوصها الإنترنت بعد حذف مكروراتها وتنقيحها باستعمال العديد من الأدوات. وتتكون هذه المدونة الضخمة من 5,8 بليون كلمة عربية في مختلف الموضوعات والأزمنة وبكل مستويات العربية ولهجاتها. ويشير الباحثون في هذا العمل إلى أنهم يعملون على تقطيع، وتجذيع 'lemmatizing' جزء من كلمات هذه المدونة الكلمات، ووسمها بأقسام الكلام بأداة 'MADA' مدى (النسخة 2.3). وهذا العمل لم يكتمل توسيمه وما تم فيه وسم بنظام مدى 'MADA'. ونظام مدى مبني على نتائج البنك الشجري العربي لجامعة بنسلفانيا. وهو المشروع المحصور في مجال النصوص الصحفية في فترة زمنية محددة (Yonatan, et al., 2013, p. 13). وهذه المدونة وإن كانت ضخمة وتمثل العربية بأزمنتها وأوعيتها، إلا أنها خليط أيضا من اللهجات العربية، وليست دراسة تخصصية لغوية، فضلا عن أنها لم تكتمل.

وفي سياق المدونات الموسمة آليا أيضا، أنشأ الحاج والخولي (2013) مدونة كلمات 'KALIMAT' التي تتكون من أكثر من عشرين ألف مقالة جمعت عشوائيا من صحيفة الوطن العمانية تتضمن موضوعات مختلفة شملت الثقافة، والاقتصاد، والأخبار العالمية والمحلية، والدين، والرياضة، وأكثر المقالات جاءت في الرياضة ثم الدين وأقلها في الأخبار العالمية. ويشير الباحثان في عملهم إلى أهمية المدونات في مجال معالجة اللغات الطبيعية. ومن بين المهام التي من أجلها أنشئت المدونة، كانت مهمة تجريب موسم ستانفورد 'Stanford' (Toutanova, et al., 2003) المدرَّب على البنك الشجري العربي 'ATB' ذي النصوص الصحفية، والمتكون من33 قسمًا للكلام، بالاعتماد على مشروع البنك الشجري الإنجليزي. وقد حقق في مدونتهما دقة توسيم وصلت إلى 96.5%. ويتفق عمل الباحثين مع ما أتناوله في هذه الورقة من حيث الإسهام بمدونة موسمة نحويا بأقسام الكلام، لكنه يختلف في آلية التوسيم، والخط الزمني والجغرافي والموضوعي لنصوص المدونة، فضلا عن انطلاقة العمل الحالي انطلاقة تخصصية مبنية على معرفة بالعربية وعلومها.

وفيما يتعلق بالموسمات النحوية الآلية للنصوص العربية، فقد بنيت في معظمها من نصوص من العربية الفصحى في مجال الصحافة، فهي وإن أظهرت نتائج جيدة على نصوص مدربة عليها تراوحت بين 90%-98% (انظر الجدول 1)، إلا أن ما طبق منها على القرآن مثلا، أو على نصوص من التراث أظهر نتائج سيئة. وقد قارنت الربيعة وآخرون (2014, pp.27-36) بين نظامي 'MADA' مدى والخليل المدربين على نصوص عربية فصيحة في مجال الصحافة؛ من أجل توسيم مدونة الذخيرة 'KSUCCA' (Alrabiah, 2015) - ذات النصوص التراثية - بأحدهما، فأسفرت النتائج عن انخفاض في دقتهما بلغ من 10% إلى 15% بالتجريب على خمس عينات من أوعية مختلفة من مدونة الذخيرة. وركزت دراسة العصيمي وآخرون (2017, pp.1-

26) على تقييم المحللات الصرفية والموسمات النحوية المتاحة للأغراض البحثية والمصممة للغة العربية الفصحى الحديثة أو القديمة، وقارنت بينها. ولوحظ انخفاض في الدقة والتغطية عند تطبيق المحللات الصرفية، والموسمات النحوية على عربية التراث. أما محمد عماد (2018, pp.1-13) فقد قام بتقطيع وتوسيم مدونة دينية صغيرة الحجم بأقسام الكلام بطريقة شبه آلية. وهي مدونة تتكون من الكتيبات التالية: الأحاديث النووية ومتن أبي شجاع والمنقذ من الضلال. وتضمنت 9000 كلمة تقريبا. ودرب أداته عليها، بالاعتماد على نفس مجموعة وسوم البنك الشجري العربي. وأظهرت النتائج أن أداته في توسيم النصوص العربية التراثية تفوقت على الأداة التي دربت على البنك الشجري العربي على الرغم من أن البنك الشجري كان أكبر بإحدى وعشرين مرة. واستكشف أبو زينة وعبد الباسط (2019) أداء موسم ستانفورد لأقسام الكلام على عربية التراث، وقيمه في القرآن الكريم على أفعال الأمر فقط. وعلى الرغم من الدقة التي رصدت عند بناء نظام ستانفورد والتي بلغت 96.26% لكل الوسوم، حقق الموسم في أدائه على أفعال الأمر في القرآن دقة منخفضة جدا بلغت 7.28%. ونظرت التميمي (2020، صص. 120-166) في واقع المجموعات التوسيمية النحوية لكل موسم من الموسمات النحوية الثلاثة (ستانفورد-مدى–أميرا) خلال مدونات طبقت عليها هذه الأنظمة ولم تدرب عليها. وبينت أداءها اللغوي والتقني من خلال أول خمسين كلمة فقط من الكلمات الأكثر تكراراً في كل وسم لكل مجموعة. وتضمن الأداء اللغوي النظر في صلاحية كل مجموعة وسوم للعربية ثم النظر في محتوى وسومها للتأكد من مطابقة مسمى الوسم لمحتواه، من خلال تطبيقها على المدونات. وفي الأداء التقني، نظرت التميمي في تقييمها للموسمات إلى جانب الصحة ‘accuracy’ في المخرجات (الخمسين كلمة الأولى الأكثر تكراراً في كل وسم)، لمعرفة الكلمات الموسمة توسيما صحيحاً وتحققت من ذلك بالفحص اليدوي، وبالنظر في سياقاتها. وأسفر بحثها عن أن هذه الموسمات لم تنجح في تطبيقها على هذه المدونات؛ لأنها لم تراع قواعد العربية في مجموعة الوسوم شكلا ومضمونا، ولم تُدرّب على نصوص متنوعة تغطي عصورا للعربية قديمة وحديثة، فحققت نسبة الصحة في هذه الأنظمة على التوالي: 44,68% - 32,32%- 27,75%.

جدول (1) دقة الموسمات النحوية العربية على النصوص التي دربت عليها

الموسم	الصحة
موسم خوجة (Khoja, et al., 2001)	%90
موسم كنعان (Kanaan, et. al, 2003)	%93
موسم ستانفورد (Toutanova, et al., 2003)	%96,5
موسم أميرا (Diab, et al., 2004)	%95,49
الموسم المعتمد على الخصائص الصرفية الوظيفية (Hajic, et. al., 2005)	-95,25 %97,37
موسم فان دن بوش (Van den Bosch, et. al., 2007)	%91,5
الموسم المعتمد على الأوزان الصرفية (Alqrainy, et. al., 2008)	%91
موسم النصوص غير المشكّلة (Al-Taani & Al-Rub, 2009)	%94
موسم الحاج (Elhadj, 2009)	%96
الموسم النحوي المحسن من خلال التحليل الصرفي (Albared, et. al., 2011)	%96,6
الموسم النحوي بخاصيتي العدد والنوع (Darwish, et. al, 2014)	%98.1
موسم مدى (Habash, et al., 2013)	%96
موسم مداميرا (Pasha, et al. 2014)	%95,9

وحيث لم يقم أي لغوي متخصص في العربية حسب علمي بعملية توسيم نحوي لمدونة عربية منطلقا فيها من نظرية لغوية حديثة لأقسام الكلام من أجل بناء نموذج للتوسيم النحوي، فإن أهمية هذا العمل تكمن في كونه امتدادا لأول مشروع لغوي عربي يتصدى فيه باحث متخصص في اللغة العربية للقيام بهذه المهمة (التميمي، 2020). ويهدف إلى جمع مدونة لغوية شاملة ومتوازنة تتخذ من الإطار النموذجي للمدونة العربية (المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية) 'KACSTAC' إطارا لها. جُمعت 10 آلاف كلمة وقطعت يدويا وفق قائمة من متغيرات التقطيع التي يعتمد عليها التوسيم النحوي الهدف؛ ثم وسمت يدويا بثلاث مجموعات توسيمية نحوية مقترحة، منطلقة من قواعد النحو العربي، وصالحة للتوسيم الآلي، ويمكن الاعتماد عليها في بناء أنظمة للوسم الآلي التي بدورها يمكن أن تستعمل في وسم مدونات أخرى. وانتهى المشروع ببناء نظام آلي للتوسيم النحوي قائم على ثلاث نماذج 'CRF' أولية للتوسيم النحوي مدربة على مدونة المشروع: نموذج 'CRF' للتوسيم النحوي بمجموعة وسوم التميمي الأساسية، ونموذج 'CRF' للتوسيم النحوي بمجموعة وسوم التميمي الفرعية، ونموذج 'CRF' للتوسيم النحوي بمجموعة وسوم التميمي الموسعة، وقد ظهرت نتائج الصحة فيها على التوالي: 91,5%، 82%، 72,1%. ثم زيدت مدونة المشروع 500 كلمة، وأظهرت النتائج تحسنا طفيفا جدا يشير إلى إمكانية زيادة حجم مدونة المشروع لتحسين أداء النماذج.

3 الأدوات والبيانات

الأدوات التي سنحتاجها لإعادة تدريب النموذج الأولي:

1.3. مجموعة وسوم التميمي الأساسية

تنطلق مجموعة وسوم التميمي الأساسية (التميمي، 2020، ص. 183) من نظرية تمام حسان في تقسيم الكلام العربي (حسان، 1994، صص. 90- 120) وهو تقسيم منطقي قابل للتمثيل الهيكلي ويمكن من خلاله استقراء جوانب الكلمات، واستيعابها بحيث لا تفلت أي منها من أيّ قسم من أقسام الكلمة، فضلا عن إمكانية تكييفه مع تعليمات المجموعة الاستشارية الخبيرة لمعايير هندسة اللغة 'EAGLES' الساعية لتوحيد معايير شكل ومحتوى التوسيم النحوي (McEnery & Wilson, 2011, p. 38). وتأتي مجموعة التميمي الأساسية في المستوى الأول الذي حددته 'EAGLES' على سبعة أقسام مع إمكانية توسيعها، وهي: الأسماء - الأفعال - الصفات - الضمائر - الظروف - الأدوات - الخوالف، بالإضافة إلى التقسيمات غير اللغوية التي يستلزمها التحليل النصي الحاسوبي، وهي: علامات الترقيم-الاختصارات-الكلمات الأجنبية-الرموز. ويظهر الجدول (2) الأقسام الكلامية الأساسية مع وسمها، بالإضافة إلى الأقسام غير اللغوية ووسومها.

جدول (2) الأقسام الكلامية الأساسية مع وسمها

القسم الكلامي الأساسي	وسمه
الاسم	N
الفعل	V
الصفة	A
الضمير	P
الأداة	RP
الخالفة	I
الظروف	D
علامات الترقيم	PUNC
الرموز	SYMB
ألفاظ أجنبية	FOREIGN
اختصارات	ABBREV
أرقام	DIGIT

2.3. المدونة الفرعية الموسمة يدويا على المستوى النحوي

وهي مدونة المشروع السابق (التميمي، 2020، صص. 168-175)، وقد جمعت هذه المدونة الفرعية لغرض بناء مقطع وموسم نحوي يلبي حاجة الدراسات العربية المعتمدة على المدونات ويقبله المتخصصون في العربية، بحيث يكون مبنيا على المعرفة بنحو العربية. فبنيت المدونة بنفس خصائص المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية المعروفة بالمدونة العربية، حيث تضم عشرة آلاف كلمة من النصوص العربية الفصيحة، وتتخذ من إطارها إطارا نموذجيا تراعي فيه محتوى الأوعية نسبة وتناسبا، فضلا عن مراعاتها للخط الزمني والجغرافي. وقد عولجت نصوص هذه المدونة على مراحل بدءا بتفريق كلماتها عن علامات الترقيم والرموز والأرقام وإزالة التشكيل وما إلى ذلك، ثم فصلت متغيرات التقطيع التي تستدعيها مجموعة وسوم التميمي النحوية عن الكلمات، كفصل واو العطف وباء الجر ولام الأمر وغيرها، ثم التسوية الهجائية التي تستدعيها طبيعة الكلمات العربية، كتصحيح ال التعريف التي تحذف ألفها إذا اتصلت باللام وكانت لامها شمسية، نحو: (ل لشمس) وغير ذلك، إلى أن ننتهي إلى إسناد الوسم النحوي المناسب لكل كلمة. وقد جرت فيها عملتي التقطيع والتوسيم يدويا.

3.3. نموذج 'CRF' الأولي للتوسيم النحوي

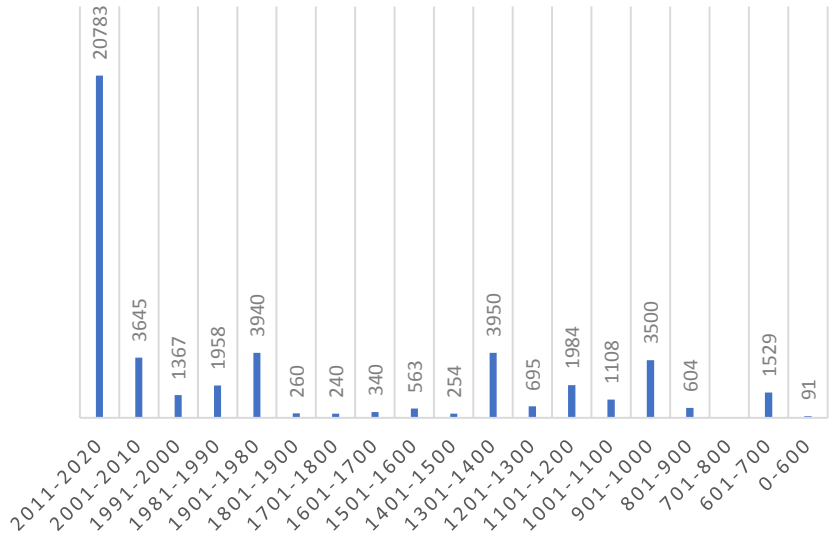
وهو نتاج خوارزمية 'CRF' بعد تدريبها على مدونة التدريب التي تشكل 80% من مدونة المشروع السابق الفرعية ذات العشرة آلاف كلمة، ثم تجريبه على مدونة الاختبار التي تشكل المتبقي من المدونة، وهو 20% من كامل المدونة. واستعنت في ذلك بلغة البرمجة (البايثون Python)، واستعملت مكتبة 'sklearn_crfsuite' لتدريب الخوارزمية بالإعدادات الافتراضية لها، بالإضافة إلى الدالة 'features' لاستخلاص بعض الخصائص المتعلقة بالتوسيم بعد أن جهزت بيانات التدريب والاختبار. وكانت الخصائص من خلال كل كلمة داخل جملة، وهي كالتالي:

- الكلمة نفسها والكلمة السابقة واللاحقة لها.
- الكلمة نفسها وحروفها الثلاثة الأولى والأخيرة.
- ما إذا كانت الكلمة رقما أم لا.
- ما إذا كانت الكلمة في أول جملة أو نهايتها.

ثم قيس الأداء في النموذج، وبلغت نسبة الصحة 91,58% (التميمي، 2020، ص. 238-243).

4.3. المدونة الإضافية الخام

وهي عبارة عن نصوص إضافية جمعتها لتكون امتداد المدونة المشروع السابق الفرعية، ذات العشرة آلاف كلمة. وتتكون من 46,811 كلمة فعلية جمعت من النصوص العربية الفصيحة على امتداد عصورها وأمكنتها، كما في الشكلين (1) و(2)، ووفقا للإطار النموذجي للمدونة العربية في الجدول (3).



شكل (1) نسبة توزيع النصوص على الفترة الزمنية في المدونة الإضافية

جدول (3) عدد الكلمات في المدونة الإضافية في كل وعاء

الأوعية	عدد الكلمات الفعلية
الصحف	9126
المخطوطات المحققة	5405
الكتب	7082
المجلات	4504
الرسائل الجامعية	5140
الدوريات المحكمة	3310
الإنترنت	4544
المناهج الدراسية	3700
وكالات الأنباء	2000
الإصدارات الرسمية	2000
المجموع	46,811

شكل (2) التوزيع الجغرافي لنصوص المدونة

5.3. مصفوفة الإرباك ‘confusion matrix’

لقياس أداء النموذج المحسن سنستعمل مدونة الاختبار التي ستشكل 20% من المدونة الكبرى (مدونة العشرة آلاف كلمة + المدونة الإضافية) بنسختيها: النسخة الخام والنسخة الموسمة نحويا. ومن المفترض أن يتنبأ النموذج بالوسوم كلها بطريقة صحيحة. ولكن في معظم الأحوال تكون بعض الوسوم المتنبئ بها لبعض الكلمات إما خاطئة وتصنف (FP) ‘false positive’ أو صحيحة وتصنف (TP) true positive، ويكون النموذج في تنبؤاته الصحيحة قد استبعد وسوما أخرى وتصنف (FN) ‘false negative’، أو يكون قد استبعد في استجاباته الخاطئة وسما صحيحا فيكون (TN) ‘true negative’. وهكذا، يُقيّم النموذج حسب هذه التصنيفات بمقارنة مدونة الاختبار الخام بمدونة الاختبار الموسومة يدويا فيما يعرف بمصفوفة الارتباك‘confusion matrix’ أو مصفوفة الخطأ ‘error matrix’، ثم يقيس المتوسط التوافقي ‘f-measure’ بين الدقة والاسترجاع المستعمل لقياس الأداء الكلي، والملائم في الحالة التي تكون أخطاء النموذج فيها من نوع FP أو FN، وأيضا حين يراد قياس عدد الأخطاء بالنسبة لعدد الاستجابات الصحيحة TP، وحين تكون الوسوم من نوع TNغير مهمة، ولا يمكن الحصول على درجته إذا كانت درجة الدقة أو درجة الاسترجاع منخفضتين جدا (Power, 2011, pp. 37-63). وتَحسِب هذه المصفوفة مما سبق القيم التالية:

1) الصحة accuracy: $= \frac{TP+TN}{TP+TN+FP+FN}$

2) الدقة precision: $= \frac{TP}{TP+FP}$

3) الاسترجاع recall: $= \frac{TP}{TP+FN}$

4) مقياس ف F-measure $= \frac{Precision.Recall}{Precision+Recall}.2$

4 المعالجة

عولجت المدونة الإضافية الخام بنفس الطريقة التي جهزت فيها مدونة العمل السابق ذات العشرة آلاف كلمة؛ تمهيدا لتطبيق إجراءات تحسين نموذج 'CRF' الأولي بها. فمرت بالمراحل التالية:

1.4. التفريق

وفيها استعنت بمعالج الكلمات 'WORD' لتعديل المسافات المضاعفة لمسافة واحدة، وبالمشذب العربي لفصل الرموز بأشكالها (@#$%^&*><×÷) وعلامات الترقيم بأنواعها (. ! ؟ , " : ؛ ، - ( )~ [ ]) والأعداد الرقمية المفردة والمركبة (12345...- 45 48 ...) عن الكلمات التي تتصل بها.

2.4. التسوية الهجائية

وهي على مرحلتين:

أ- التسوية قبل التقطيع: أزيلت فيها علامات التشكيل، والكشيدة عن الكلمات، وسوّيت الأرقام بتحويلها من العربية للهندية. ونفذت في نفس المرحلة التي فصلت فيها الكلمات عن الأرقام والرموز بالاستعانة ببرنامج المشذب العربي أيضا.

ب- التسوية بعد التقطيع: ظهرت الحاجة إليها في هذه المرحلة بسبب طبيعة الكلمات العربية التي يتغير شكلها الكتابي بعد فصلها. فكلمة مثل: ممّن، أصبحت بعد التقطيع (م من) وكان لا بد من إعادة (م) لشكلها الطبيعي قبل إدغامها في كلمة (من) لتكون كلمتين منفصلتين ومستقلتين، على النحو التالي: (مِنْ مَنْ)، وكلمات أخرى أيضا، كالكلمات المعرفة بأل التي تتصل بها الأداة (ل)، نحو: للأيام وحذفت ألفها.

3.4. التقطيع

في هذه المرحلة بدأ العمل يدويا على فصل متغيرات التقطيع. وقد حددت متغيرات التقطيع التي تقتضيها عملية التوسيم النحوي كما يظهر في الجدول (4). ولأن التوسيم النحوي يقضي بأن يكون كل متغير من متغيرات التقطيع عبارة عن مبنى تقسيمي، فما قررت عدم فصله هو مبان تصريفية لا تقسيمية. والمباني التقسيمية متميزة دائما بالصدارة في الكلمة.

جدول (4) متغيرات التقطيع التي تقتضيها عملية التوسيم النحوي

المتغير	مثاله
همزة الاستفهام والتسوية (أ)	أيرضيك ذلك؟ - "وسواء عليهم أأنذرتهم أم لم تنذرهم"
حرف الجر والقسم (ب)	ببيتك -بالله
تاء القسم (ت)	تالله
حرف الاستقبال (س)	سيزور
عن المدغمة نونها في ميم من وما (ع)	عمّن – عمّا -عم
فاء الشرطية والسببية والاستئنافية	إذا مررت فسلمْ عليهم- ولا يؤذن لهم فيعتذرون-

والعاطفة (ف)	فتعالى الله عما يشركون- دخل محمد فخالد
حرف الجر (ك)	كالقمر
حرف الجر ولام الأمر ولام التأكيد (ل)	للمنزل-لينته عنه-إن في ذلك لعبرة
من المدغمة نونها في ميم من وما (م)	ممن-مما
هاء التنبيه (ها – هـ)	أيها- هأنت
حرف العطف والمعية والحال والقسم (و)	كبر وعظم- سرت والليل-مضيت وأنا سعيد -والله
من الجارة المدغمة ميمها في ميم الاستفهام (م)	ممّ
في الجارة المتصلة بمن (في)	فيمن
حروف الجر المتصلة بما الاستفهامية المحذوفة الألف (ب-في-ل- على-إلى-حتى)	بم –فيم-لم-علام-إلام-حتام
إذ الظرفية المتصلة ببعض الكلمات	حينئذ-عندئذ

4.4. توسيم المدونة الإضافية الخام بنموذج 'CRF' الأولي

وسمت المدونة الخام بنموذج 'CRF' الأولي على مراحل، وذلك باستعمال منهج إعادة تدريب النموذج. ففي المرحلة الأولى جرب النموذج على خمسة آلاف كلمة من المدونة الإضافية الخام المفرقة والمقطعة يدويا لتوسيمها نحويا بالوسوم الأساسية، ثم بعد تجريب النموذج الأولي عليها وظهور النتائج، صححت الأخطاء فيها يدويا، وضمت تلك الخمسة آلاف كلمة الموسومة نحويا والمصححة يدويا للمدونة الفرعية التي بني منها النموذج الأولي في العمل السابق، حيث وزعت بين مدونتي التدريب والاختبار فيها، فكانت 4000 كلمة لمدونة التدريب و1000 لمدونة الاختبار. أعيد تدريب نموذج 'CRF' بالكامل على مدونة التدريب بعد زيادة حجمها وزادت دقته قليلا. وبعد أن أصبح لدينا نموذجا محسنا أعيد العمل بالتوسيم النحوي من خلاله على خمسة آلاف كلمة أخرى من المدونة الخام مع تصحيح المخرج يدويا، وإضافته لنصوص المدونة الفرعية في العمل السابق، وإعادة تدريب النموذج من جديد على ما سبق. وهكذا، حتى تم توسيم 46,811 كلمة بإعادة تدريب النموذج أكثر من مرة، ثم إضافة الكلمات للمدونة الفرعية في العمل السابق.

5.4. إعادة تدريب النموذج واختباره للمرة الأخيرة

بعد أن أصبح لدينا مدونة موسمة نحويا من 60,919 كلمة فعلية (المدونة السابقة مع المدونة الحالية)، درب نموذج 'CRF' للتوسيم النحوي بمجموعة وسوم التميمي الأساسية على 80% من المجموع الكلي للمدونتين (48,979 كلمة فعلية) للمرة الأخيرة. ثم اختبر على 10% وهو المتبقي منها (11,940 كلمة فعلية)، وذلك للتقييم النهائي للنسخة المحسنة من النموذج.

5 النتائج

لقد حقق نموذج 'CRF' في نسخته الأولية المدربة على عشرة آلاف كلمة درجة صحة بلغت 91,58%. وبعد إعادة تدريبه بالكامل على 80% من المدونة الفرعية الجديدة المكونة من 48,979 كلمة فعلية، والموسمة شبه يدويا بإعادة تدريب النموذج، جرب النموذج على ال 20%

المتبقية منها. وقد تحسن أداؤه عن النسخة الأولية منه، وحقق درجة صحة بلغت 93.97%. إلا أن قيم الأداء انخفضت في الوسوم اللغوية ذات الأمثلة الأقل، كالظروف (D)، والخوالف (I) كما يظهر في الجدول (5). ويلاحظ أيضا في الجدول (5) أن الأداء في الكلمات الوظيفية كالضمائر والأدوات جاء أعلى من الأداء في كلمات المحتوى كالأفعال والأسماء والصفات حيث إنها من القلة الأكثر تكرارا في المدونة على عكس كلمات المحتوى التي قد لا ترد في المدونة سوى مرة واحدة.

جدول (5) قيم الأداء في النموذج المحسن

الوسوم	الدقة	الاسترجاع	مقياس ف	عدد الأمثلة
N	0.914	0.951	0.932	4423
RP	0.983	0.976	0.979	3397
PUNC	0.998	0.995	0.997	1286
V	0.932	0.889	0.910	1083
A	0.828	0.791	0.809	1035
P	0.958	0.914	0.935	444
DIGIT	0.991	0.991	0.991	114
D	0.882	0.909	0.896	33
I	1.000	0.438	0.609	16
FOREIGN	1.000	0.784	0.879	74
ABBREV	0.875	0.700	0.778	30
SYMB	1.000	1.000	1.000	5
المقاييس العامة: الصحة = 0.9397 مقياس ف= 0.9394 الدقة= 0.9399 الاسترجاع = 0.9398				

ويظهر التحسن واضحا في التوسيم النحوي بتطبيق النموذجين (الأولي والمحسن) على نفس المقتطفات من النصوص المتنوعة التي لم يدرب عليها النموذجين. ونرى مثلا كيف تمكن النموذج المحسن من تمييزه للأفعال والصفات عن الأسماء، كما يلي:

1) **نص قرآني كريم (سورة الأعراف):** (... يا قوم لقد أبلغتكم رسالة ربي ونصحت لكم ولكن لا تحبون الناصحين)

- النص موسم بالنموذج الأولي:

('يا', 'RP'), ('قوم', 'N'), ('ل', 'RP'), ('قد', 'RP'), ('أبلغتكم', 'V'), ('رسالة', 'N'), ('ربي', 'N'), ('و', 'RP'), ('نصحت', 'V'), ('لكم', 'RP'), ('و', 'RP'), ('لكن', 'RP'), ('لا', 'RP'), *('تحبون', 'N'), ('الناصحين', 'N')*

- النص موسم بالنموذج المحسن:

('يا', 'RP'), ('قوم', 'N'), ('ل', 'RP'), ('قد', 'RP'), ('أبلغتكم', 'V'), ('رسالة', 'N'), ('ربي', 'N'), ('و', 'RP'), ('نصحت', 'V'), ('لكم', 'RP'), ('و', 'RP'), ('لكن', 'RP'), ('لا', 'RP'), *('تحبون', 'V'), ('الناصحين', 'A')*

2) **نص تراثي للقرطبي (520هـ):** "وقد كان بعض الأصحاب سألني أن أمهد في أول كل كتاب منه مقدمة تنبئ عليه مسائله من الكتاب والسنة، وترد إليها بالقياس عليها مع الربط لها بالتقسيم والتحصيل لمعانيها".

- النص موسم بالنموذج الأولي:

('و', 'RP'), ('قد', 'RP'), ('كان', 'RP'), ('بعض', 'N'), *('الأصحاب', 'N'), ('سألني', 'N'),* ('أن', 'RP'), ('أمهد', 'N'), ('في', 'RP'), ('أول', 'A'), ('كل', 'N'), ('كتاب', 'N'), ('منه', 'RP'), ('مقدمة', 'A'), ('تنبئ', 'V'), ('عليه', 'RP'), ('مسائله', 'N'), ('من', 'RP'), ('الكتاب', 'N'), ('و', 'RP'), ('السنة', 'N'), ('،', 'PUNC'), ('و', 'RP'), ('ترد', 'V'), ('إليها', 'RP'), ('ب', 'RP'), ('القياس', 'N'), ('عليها', 'RP'), ('مع', 'N'), ('الربط', 'N'), ('لها', 'RP'), ('ب', 'RP'), ('التقسيم', 'N'), ('و', 'RP'), ('التحصيل', 'N'), ('ل', 'RP'), ('معانيها', 'N')

- النص موسم بالنموذج المحسن:

('و', 'RP'), ('قد', 'RP'), ('كان', 'RP'), ('بعض', 'N'), *('الأصحاب', 'A'), ('سألني', 'V'),* ('أن', 'RP'), ('أمهد', 'V'), ('في', 'RP'), ('أول', 'A'), ('كل', 'N'), ('كتاب', 'N'), ('منه', 'RP'), ('مقدمة', 'A'), ('تنبئ', 'V'), ('عليه', 'RP'), ('مسائله', 'N'), ('من', 'RP'), ('الكتاب', 'N'), ('و', 'RP'), ('السنة', 'N'), ('،', 'PUNC'), ('و', 'RP'), ('ترد', 'V'), ('إليها', 'RP'), ('ب', 'RP'), ('القياس', 'N'), ('عليها', 'RP'), ('مع', 'N'), ('الربط', 'N'), ('لها', 'RP'), ('ب', 'RP'), ('التقسيم', 'N'), ('و', 'RP'), ('التحصيل', 'N'), ('ل', 'RP'), ('معانيها', 'N')

3) **نص معاصر من مقال للمحيفيظ في صحيفة المغرب (2020):** "لنركز على التصدي لخطابات الكراهية والتمييز العنصري، التي تتزايد بشكل مقلق وتكتسب جرأة ووقاحة لم تكن تجرؤ عليها إلى وقت قريب".

- النص موسم بالنموذج الأولي:

('ل', 'RP'), ('نركز', 'N'), ('على', 'RP'), ('التصدي', 'N'), ('ل', 'RP'), ('خطابات', 'N'), ('الكراهية', 'N'), ('و', 'RP'), ('التمييز', 'N'), ('العنصري', 'N'), ('،', 'PUNC'), ('التي', 'P'), ('تتزايد', 'V'), ('ب', 'RP'), ('شكل', 'N'), *('مقلق', 'N'),* ('و', 'RP'), ('تكتسب', 'V'), ('جرأة', 'N'), ('و', 'RP'), ('وقاحة', 'N'), ('لم', 'RP'), ('تكن', 'RP'), *('تجرؤ', 'N'),* ('عليها', 'RP'), ('إلى', 'RP'), ('وقت', 'N'), ('قريب', 'A')

- النص موسم بالنموذج المحسن:

('ل', 'RP'), ('نركز', 'N'), ('على', 'RP'), ('التصدي', 'N'), ('ل', 'RP'), ('خطابات', 'N'), ('الكراهية', 'N'), ('و', 'RP'), ('التمييز', 'N'), ('العنصري', 'N'), ('،', 'PUNC'), ('التي', 'P'), ('تتزايد', 'V'), ('ب', 'RP'), ('شكل', 'N'), *('مقلق', 'A'),* ('و', 'RP'), ('تكتسب', 'V'), ('جرأة', 'N'), ('و', 'RP'), ('وقاحة', 'N'), ('لم', 'RP'), ('تكن', 'RP'), *('تجرؤ', 'V'),* ('عليها', 'RP'), ('إلى', 'RP'), ('وقت', 'N'), ('قريب', 'A')

وبنظرة في مدونة الاختبار التي تكونت من 11940 كلمة فعلية، وقفت على جميع التنبؤات الخاطئة للنموذج المحسن، ووجدتها 633 خطأ صنفتها إلى ستة أقسام، كما يظهر في الجدول (6).

جدول (6) التكرار لأنواع الأخطاء في الموسم النحوي المحسن

نوع الخطأ	نسبة التكرار
1- كلمات غير معروفة	%52.60
2- اللبس اللغوي	%17.53
3- عدم الاتساق في مدونة التدريب	%14.21
4- أخطاء في مدونة الاختبار	%5.21
5- الفجوة المعجمية	%9.79
6- أخطاء كتابية	%0.63

ويظهر أن ما يشكل نصف أخطاء النموذج تقريبا يقع في كلمات غير معروفة ‘unknown word’. وأقصد بالكلمات غير المعروفة هي تلك التي وردت في مدونة الاختبار، ولم ترد في مدونة التدريب. ومن المفترض هنا أن النموذج يعتمد على السياق في تعيينه لنوع الكلمة، ولذا ليس من الغريب أن نجده يحدد نوع الكلمة (الآخذ) في المثال التالي تحديدا دقيقا، وإن لم تكن قد وردت في مدونة التدريب:

و_RP هو_P النهر_N الآخذ_A من_RP الفرات_N

أما أخطاء اللبس اللغوي فالمقصود بها تلك الأخطاء التي تقع بسبب اللبس الخطابي أو الدلالي أو الصرفي، نحو الأخطاء في (نِعْم ونعَم) و(ذِكْر وذَكَرَ) و(ذا الاسم وذا الضمير الإشاري) و(جَهِلَهُ وجَهْلُهُ). وتحدث أخطاء قليلة أخرى لا تشكل سوى 0.63% من أخطاء مدونة الاختبار، قد تدخل في اللبس لكن اللبس فيها يقع على المستوى الكتابي، ولم يحدث ذلك إلا في ثلاث كلمات هي: (الذي–على – اول).

وأعني بأخطاء عدم الاتساق هي الأخطاء التي تقع من الموسِّم البشري في مدونة التدريب، فلا يكون هناك انتظام في توسيم الكلمة فتجدها في مدونة التدريب تارة بوسم الاسم وتارة بوسم الصفة. وقد وقع ذلك في 14.21 كلمة في مدونة الاختبار. وقد وسم النموذج منها 5.37% بالوسم الصحيح، وهذا يعني أن النموذج صحح الأخطاء التي وقع فيها الموسم البشري. ومن هذه الأخطاء التي صححها النموذج: (مقارنة – شوارع – الكائنات – وسيلة).

وثمة أخطاء وقعت في مدونة الاختبار تسببت بظهور توسيمات عُدت خطأ، رغم أن النموذج قد أصاب في تعيينها، بسبب خطأ من الموسم البشري في مدونة الاختبار. منها مثلا:

يؤسس_V توازن_N *بين_D* الاستقلال_N و_RP *الحرية_A* (النص في مدونة الاختبار بالتوسيم البشري)

يؤسس_V توازن_N *بين_N* الاستقلال_N و_RP *الحرية_N* (النص المتنبئ به النموذج)

وأخيرا نرى أن الأخطاء التي صنفتها تحت الفجوة المعجمية. ويقصد بالفجوة المعجمية ‘lexical gap’ ما عرفه مانق (2011, p.176) بالأخطاء التي تحدث في الكلمة الواردة مرارا في مدونة التدريب بوسم معين، ثم يسمها النموذج في نص آخر بوسم مخالف. أي أن الكلمة ترد في مدونة التدريب أكثر من مرة بوسمها الصحيح، ولكن لا ترد بالسياق نفسه في مدونة الاختبار فيقع الخطأ. وقد شكلت هذه الأخطاء ما يقارب 10% من مجموع الأخطاء. فقد أخطأ النموذج في توسيم كلمة (الداء) ووسمها بوصفها (صفة). وقد وردت في مدونة التدريب بالسياق:

السيد_A رئيس_A الحكومة_N شخص_V الداء_N

وفي مدونة الاختبار وردت في سياق مختلف:

يا_RP رب_N من_RP من_P الداء_A ؟_PUNC

6 الخلاصة

إن زيادة حجم مدونة التدريب لا يفيد تماما في التخلص من معظم الأخطاء التي يقع فيها النموذج، وإن كان حلا لبعض المشكلات التي تعتريه. فنجاح الموسم المحسن في توسيم كلمات معينة بدقة أكثر من توسيم النموذج الأولي، لا يعني على أية حال أن تلك الكلمات وردت في المدونة الإضافية، ولم ترد في الأولى. إذ بالرجوع للمدونتين لم يعثر مثلا على بعض الكلمات الواردة في النصوص السابقة (انظر المبحث 5)، نحو: (تحبون، الناصحين، سألني، أمهد). ومع ذلك، تمكن النموذج من تصنيفها بدقة.

وعلى أية حال، يعيننا تصنيف أخطاء النموذج في مدونة الاختبار على إمكانية التقليل منها. وبالنظر إلى الأخطاء الناتجة من عدم الاتساق في مدونة التدريب، وأخطاء مدونة الاختبار، والأخطاء الكتابية التي تشكل في مجموعها 20% تقريبا من الأخطاء، نجدها أيسر أنواع الأخطاء معالجة، إذ تعالج فقط بتصحيح الأخطاء في مدونة التدريب والاختبار. ورغم أن مناهج تعلم الآلة شبه الموجه 'unsupervised' نحو: فئات التشابه التوزيعي 'distributional similarity' classes'، مفيدة جدا في معالجة أخطاء الكلمات غير المعروفة والأخطاء الناتجة من الفجوة المعجمية، إلا أنها لا تحدث إلا تحسنا طفيفا (Manning, 2011, p.176).أما أخطاء اللبس اللغوي، فقد يفيد فيها ربط النموذج: بالتشكيل وحسابات التصاحب 'correlation'.

7 الأعمال المستقبلية

بالإضافة إلى أن توسيم مدونة ضخمة توسيما يدويا؛ لبناء نموذج لغوي صالح للتعميم على اللغة أمر صعب التنفيذ، فإن طبيعة البيانات اللغوية تتغير مع مرور الزمن ولا يمكن التحكم فيها أو تقييدها. وفي هذا الظرف تتألق منهجية إعادة تدريب النموذج 'model retraining'. وباستمرار استعمالنا لهذه الطريقة سنتيح استعمال مدونة موسمة نحويا ذات حجم يتناسب مع امتداد العربية الفصحى زمنيا وجغرافيا وموضوعيا، فيحقق أهدافا متعددة في مجالات لغوية، ولغوية حاسوبية، وأخرى غير لغوية. فضلا عن نموذج للتوسيم النحوي العربي يمكن تعميم نتائجه على نصوص العربية الفصحى في كل آن ومكان، وقابل للتحسين والتكييف بمنهجيات التحسين الأخرى في تعلم الآلة، كضبط الخوارزمية 'algorithm tuning'، وهندسة الخصائص 'features engineering'، أو استعمال منهج التجميع 'ensemble method'، وغير ذلك من منهجيات التحسين التي لا تتطلب زيادة في حجم نصوص المدونة وتوسيمها. وأخيرا، قد تكون البصيرة اللغوية المستنيرة في معالجة اللغات الطبيعية أكثر فاعلية في تحسين الأداء من التقنيات التحسينية الأخرى.

المراجع

المراجع العربية:

التميمي، أفراح. (2020). نظام آلي للتقطيع والتوسيم النحوي العربي. ط1. دار كنوز المعرفة. عمان. الأردن.

الثبيتي، عبد المحسن والطوالة، ندى والعتيبي، سعد والمرشدي، بشاير. (2012). طريقة تعتمد على المدونات اللغوية لتجهيز بيانات تدريب واختبار أنظمة الوسوم النحوية. في المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية في دورته الثامنة. القاهرة. مصر.

الحاج، يحيى. (2013). إعداد وتجهيز نظام إحصائي للتعرف الآلي على المفردات القرآنية: الخصائص والسمات الصرف-نحوية وآلية مستحدثة لوسمها. الدورة التاسعة للمؤتمر الدولي لعلوم وهندسة الحاسوب. الحمامات. تونس. حبش، نزار. (2014). مقدمة في المعالجة الطبيعية للغة العربية. ترجمة: هند الخليفة، ط1. جامعة الملك سعود. الرياض. السعودية.

حسان، تمام. (1994). اللغة العربية، معناها ومبناها. د.ط. دار الثقافة. الدار البيضاء. المغرب.

المراجع الأجنبية:

Albared, M., Omar, N., & Ab Aziz, M. J. (2011). "Improving Arabic part-of-speech tagging through morphological analysis". In Intelligent Information and Database Systems. Springer Berlin Heidelberg. pp. 317-326

AbuZeina, D., Mostafa, T., & Abdalbaset, M. (2019). "Exploring the Performance of Tagging for the Classical and the Modern Standard Arabic". Advances in Fuzzy Systems

Alosaimy, A., & Atwell, E. (2017). "Tagging Classical Arabic Text using Available Morphological Analyzers and Part of Speech Taggers". Journal for Language Technology and Computational Linguistics. (32)1: 1-26

Alqrainy, S., AlSerhan, H. M., & Ayesh, A. (2008). "Pattern-based algorithm for Part-of-Speech tagging Arabic text". Computer Engineering & Systems ICCES International Conference on IEEE. Cairo. Egypt. pp. 119-124

Alrabiah, M., Al-Salman, A., Atwell, E., & Alhelewh, N. (2014). "KSUCCA: A key to exploring Arabic historical linguistics". International Journal of Computational Linguistics (IJCL). 5(2): 27-36

Alrabiah, M. (2015). "Building A Distributional Semantic Model for Traditional Arabic and Investigating its Novel Applications to The Holy Qur'an". PhD dissertation. King Saud University. KSA.

Al-Taani, A., & Al-Rub, S. A. (2009). "A Rule-Based Approach for Tagging Non-Vocalized Arabic Words". International Arab Journal of Information Technology (IAJIT). Amman. Jordan. (6)3: 320-328

Diab, M., Hacioglu, K., Jurafsky, D. (2004). "Automatic tagging of Arabic text: From raw text to base phrase chunks". In Proceedings of HLT-NAACL, Association for Computational Linguistics, Boston: USA., pp. 149-152

Dukes K., & Habash N. (2010). "Morphological Annotation of Quranic Arabic". In The Language Resources and Evaluation Conference. Malta. pp. 2530-2536

Elhadj, Y. (2009). "Statistical Part-of-Speech Tagger for Traditional Arabic Texts". Journal of Computer Science. Riyadh. KSA. (5)11:794

El-Haj, M., & Koulali, R. (2013). "KALIMAT a Multipurpose Arabic Corpus". At the Second Workshop on Arabic Corpus Linguistics (WACL-2). Lancaster. UK. pp. 22-25

Habash, N., & Rambow, O. (2005). "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop". In Proceedings of (ACL'05). Michigan. USA. pp. 573-580

Habash N., & Roth R. (2009). "CATiB: The Columbia Arabic Treebank". Center for Computational Learning Systems. Columbia University. New York. USA.

Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). "Morphological Analysis and Disambiguation for Dialectal Arabic". In Proceedings of (NAACL-HLT). Georgia. USA. pp. 426–432

Hajic, J., Smrz, O., Buckwalter, T., & Jin, H. (2005). "Feature-based tagger of approximations of functional Arabic morphology". In Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT). Barcelona. Spain.

Kanaan, K., Al-Shalabi, R., & Sawalha, M. (2003). "Full automatic Arabic text tagging system". The proceedings of (ICITNS). Amman. Jordan.

Khoja, Sh. (2001). "APT: Arabic part-of-speech tagger". In Proceeding of the Student Workshop at the 2nd Meeting of the NAACL. Carnegie Mellon University. Pennsylvania. USA. pp. 20–25

Khoja, Sh., Garside, R., & Knowles, G. (2001). "A tagset for the morphosyntactic tagging of Arabic". In Proceedings of Corpus Linguistics Conference. Lancaster. UK. pp. 341-353

Klabjan, D., & Zhu, X. (2020). "Neural Network Retraining for Model Serving". https://arxiv.org/pdf/2004.14203v1.pdf

Lafferty, J., McCallum, A., & Pereira, F. (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco. USA. pp. 282–289

Maamouri, M., Bies A., Buckwalter T., & Mekki W. (2004). "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus". In NEMLAR Conference on Arabic Language Resources and Tools. Cairo. Egypt. pp. 102–109

Manning, C. D. (2011). "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?". Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg. Germany. pp. 171-189

McEnery, T., Xiao, R., & Tono, Y. (2006). "Corpus-Based Language Studies". Routledge. USA.

McEnery, T., & Wilson, A. (2011). "Corpus Linguistics (An Introduction)". Edinburgh University press. Edinburgh. UK.

Mohamed, E. (2018). "Morphological Segmentation and Part-of Speech Tagging for the Arabic Heritage". ACM Transactions on Asian and Low-Resource Language Information Processing. (17)3: 1–13

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., & Roth, R. (2014). "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic". In Proceedings of (LREC'14). (ELRA). Reykjavik. Iceland. pp. 1094-1101

Powers, David M W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. Australia. 2(1): 37–63

Sawalha, M. (2011). "Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING". PhD dissertation. School of Computing. University of Leeds. UK.

Toutanova, K., Klein D., Manning C. D., & Singer Y. (2003). "Feature-rich part-of speech tagging with a cyclic dependency network". In Proceedings of (NAACL). (1): 173-180

Van den Bosch, A., Marsi, E., & Soudi, A. (2007). "Memory-based morphological analysis and part-of-speech tagging of Arabic". In Soudi, A., van den Bosch A., & Neumann, G. (eds.). Arabic Computational Morphology. Springer. Berlin. pp. 203–219.

Yonatan, B., Nizar, H., Kilgarriff, A., Ordan, N., Roth, R., & Suchomel V. (2013). "arTenTen: a new, vast corpus for Arabic". In Proceedings Of (WACL'2) Workshop on Arabic Corpus Linguistics. UK. Lancaster. p. 20