INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

التلقيب الآلي للكلمات العربية باستخدم أداة التعلم الآلي YAMCHA

أحمد عبد الغني
abdelghany.ma@umk.edu.my
جامعة ماليزيا كلنتان

أحمد زكي بن أميرالدين
ahmadzaki@umk.edu.my
جامعة ماليزيا كلنتان

ملخص

التلقيب الآلي للكلمات العربية هو عملية وسم الكلمة العربية بقسم الكلام المناسب لها اعتمادًا على سياقها. وتعد هذه العملية خطوة أساسية في معظم تطبيقات معالجة اللغة الطبيعية (NLP) مثل التلخيص الآلي واسترجاع المعلومات والترجمة الآلية وغيرها من التطبيقات. ويهدف هذا البحث إلى تقديم ملقِّب آلي عربي معتمد على نظام إحصائي يستفيد من أنظمة تعليم الآلة. ونظام التعلم الآلي المستخدم في هذا البحث هو نظام YAMCHA (Yet Another Multipurpose CHunk Annotator)، وهو أداة مفتوحة المصدر تقوم بأداء الكثير من مهام معالجة اللغة، مثل التلقيب الصرفي الآلي للكلمات، والتعرف على أسماء الكيانات، والتحليل التركيبي للجمل، وغيرها من المهام اللغوية. ويستخدم YamCha خوارزمية في التعلم الآلي تسمى آلة المتجهات الداعمة (Support Vector Machines) التي تستخدم في تصنيف البيانات بدقة وكفاءة بالغة لأنها تستخدم جزء من البيانات في التدريب والتعلم، كما أنها تتيح تغيير مدى ونوع المعلومات اللغوية المعتمد عليها في التعلم الآلي (feature set and window-size). لذلك فالمنهجية المقترحة تستلزم كمية لا بأس بها من النصوص المحللة على مستوى أقسام الكلام من أجل تعليم وتدريب النظام عليها. وبلغ حجم المدونة المستخدمة في البحث 100.039 كلمة، وقد تم تقسيمها بنسبة 70% للتدريب و30% للاختبار فكان حجم مدونة التدريب 64.608 كلمة، وحجم مدونة الاختبار 35.431 كلمة، وبلغ عدد أقسام الكلام التي تدرّب عليها النظام وميّز بينها 48 لقبًا صرفيًا. وقد تم تدريب النظام على مدونة التدريب عدة مرات مع تغيير مدى المعلومات اللغوية المستخدمة في التدريب ثم تحليل مدونة الاختبار وتقييم النتائج من أجل الوصول لأفضل نتائج في التلقيب الآلي للكلمات العربية. وبلغت أقل نسبة خطأ 11.4%، وكانت في حالة اعتبار الكلمة السابقة في التحليل دون النظر إلى لقبها الصرفي (F:-1..0:0..).

الكلمات المفتاحية: YAMCHA ؛ التعلم الآلي ؛ آلة المتجهات الداعمة ؛ مدونة التدريب ؛ مدونة الاختبار

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

Automatic POS tagging of Arabic words using the YAMCHA machine-learning tool

ABSTRACT

The automatic POS tagging is the process of assigning the appropriate POS tag for each word in text depending on the context. This process is an essential step in most NLP applications such as automatic summarization, information retrieval, machine translation, and other applications. This research aims to present an Arabic POS tagger based on a statistical approach and machine learning systems. The machine learning system used in this paper is the YAMCHA (Yet Another Multipurpose CHunk Annotator) tool, which is an open source tool that performs many language processing tasks, such as automatic POS tagging, name entity recognition, syntax analysis, and other linguistic tasks. Yamcha uses an algorithm in machine learning called Support Vector Machines that is used to classify data with great accuracy and efficiency because it uses part of the data in training, and it also allows changing the range and type of linguistic information relied on in machine learning (feature set and window-size). Therefore, the proposed system requires a large amount of data analyzed at the level of POS in order to train it. Corpus used in this research has the size of 100,039 words, and it was divided by 70% for training and 30% for testing, so the size of the training corpus was 64,608 words, and the size of the testing corpus was 35,431 words, and the tag set used in training and testing was 48 morphological tags. The system was trained several times with changing the range of linguistic information used in training process, and then new texts were tested and evaluated in order to reach the best results in the automatic POS tagging. The lowest error rate achieved was 11.4%, when the previous word was considered in the training process without considering its POS tag (F: -1..0: 0..).

Keywords: YAMCHA ;machine learning system; Support Vector Machines; training corpus; testing corpus.

مقدمة

فهم الحاسب للغات الطبيعية من المشكلات الكبرى التى تواجه نظم المعالجة الآلية للغات الطبيعية، لأن ذلك يتطلب معرفة عميقة بالعالم الخارجى إلى جانب معرفة القواعد اللغوية جيدًا ومحاولة محاكاتها وتمثيلها رياضيًا بالشكل الذي يفهمه الحاسب الآلى، والذي يزيد من تعقيد معالجة اللغة العربية على وجه الخصوص أن الكثير من القضايا والاستخدامات اللغوية ليس لها قواعد مطَّردة وإنما تعتمد على السماع والتلقي وليس القياس والقواعد المعيارية، مثل اشتقاق المصادر الثلاثية وصيغ جمع التكسير وغير ذلك، أيضًا من المشكلات التي تواجه معالجة اللغة العربية خصوصًا مشكلة الغموض واللبس اللفظي الناتج عن عدم تمثيل علامات التشكيل في الكتابة العربية إلى جانب اللبس اللفظي الناتج عن الاشتراك في اللفظ والنطق، والموجود في جميع اللغات الطبيعية، كذلك الطبيعة التلاصقية المعقدة للكلمة العربية التي يسمح بها صرف اللغة العربية وتجعل الكلمة المفردة تتكون من وحدات صرفية (بل ونحوية) كثيرة مثل "فسيكفيكهم"، و"زوجناكها"، و"فسينفقونها"، مما يجعل الكلمة المفردة تُترجَم بجملة كاملة الأركان في اللغات الأخرى، إلى جانب ظاهرة استتار الضمائر وعدم ظهورها في السياق اللغوي وغيرها من المشاكل التي لا توجد في الكثير من اللغات الأخرى والتي تصعّب من عملية ميكنتها وتطويعها للمعالجة الآلية.

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

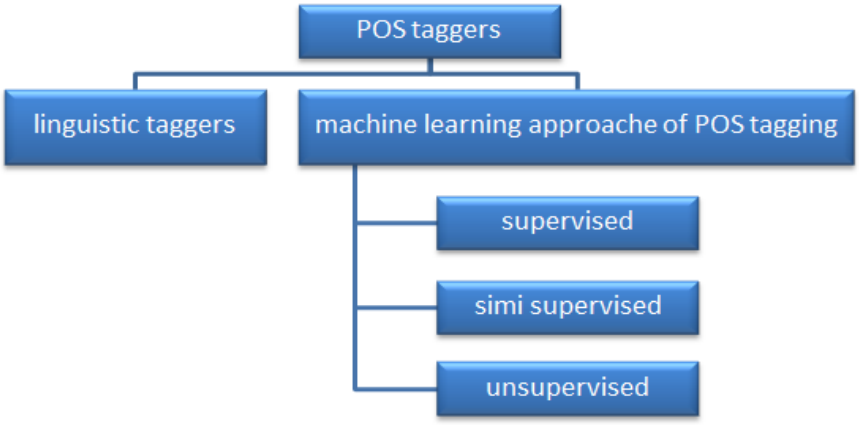
من أجل ذلك اتجهت أنظار المشتغلين بمعالجة اللغة إلى الأنظمة المعتمدة على المدونات الضخمة والأنظمة الإحصائية التي تستفيد من خوارزميات رياضية ليس لها علاقة بالقواعد اللغوية من أجل تجاوز التحديات اللغوية التي تواجه أنظمة المعالجة المعتمدة على القواعد.

والمدونة اللغوية هي مجموعة من النصوص اللغوية المحددة الحجم والمعروضة في شكل آلي، مكتوبةً كانت أو منسوخةً صوتياً (TRANSCRIBED) لكلام تم تسجيله مسبقاً تم جمعُها بطريقة مدروسة وممثلة للواقع اللغوي من أجل دراسة ظاهرة لغوية معينة أو من أجل الدراسة اللغوية بصفة عامة [1]. وقد أثبت ذلك الاتجاه في دراسة اللغة نجاحاً كبيراً وتفوقاً ملحوظاً على الاتجاه التقليدي في دراسة اللغة، فالمدونات اللغوية تعبر عن اللغة بكل تفاصيلها واستخداماتها وما هو كائن بالفعل (DESCRIPTIVE APPROACH) أما القواعد النحوية والصرفية للغة فهي تصف ما يجب أن تكون عليه اللغة (PRESCRIPTIVE APPROACH) وليس ما هو كائن بالفعل، لذلك كان منطقياً توقُّع تفوُّق الاتجاهات المعالجية التي تعتمد على المدونات على نظيراتها المعتمدة على القواعد اللغوية. والملقّب الصرفي CLAWS [2] يعتبر دليلاً واضحاً على ذلك، فقد اعتمد في بنائه على مدونة BROWN المحللة فأعطى نسبة صحة وصلت إلى 97% مقارنةً بقرينه TAGGIT [3] الذي صممه GREENE و RUBIN عام 1971 من أجل تحليل مدونة BROWN وكان معتمداً في بنائه على قواعد اللغة فأعطى نسبة صحة وصلت إلى 77% [15] .

وظهر تفعيل علم الإحصاء في مجال اللغويات الحاسوبية عام 1990 تحت مسمى تقنيات تعليم الآلة (MACHINE LEARNING TECHNIQUES)، وأدّى ذلك إلى ثورة عارمة اجتاحت معظم تطبيقات المعالجة الآلية للغة الطبيعية، فطُرِحت الأفكار والخوارزميات من أجل الاستفادة من ذلك الاتجاه الجديد في حل إشكاليات اللغة [4].

اتجاهات بناء الملقِّبات الصرفية

معالجة الالتباس الصرفي هو عملية فحص لكلمات سياق معين، وتحديد الوسم الصرفي المناسب المقصود لكل كلمة في هذا السياق. ويوجد اتجاهان رئيسان لمعالجة التباس اللقب الصرفي، هما الاتجاه المعياري المعتمد على قواعد اللغة (LINGUISTIC TAGGERS)، والاتجاه الوصفي المعتمد على المدونات وأنظمة تعليم الآلية (MACHINE-LEARNING APPROACHES) الذي يتفرع منه الأنظمة الموجَّهة في تعلم الآلة (SUPERVISED LEARNING)، والأنظمة شبه الموجَّهة (MINIMALLY AND SEMI-SUPERVISED LEARNING)، والأنظمة غير الموجَّهة (UNSUPERVISED LEARNING)، كما هو موضح في شكل (1)[5].



شكل 1. اتجاهات التلقيب الآلي للكلمات

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

أ- الاتجاه المعياري المعتمد على قواعد اللغة

يعتمد هذا النوع من الملقبات الصرفية على عدد لا حصر له من القواعد اللغوية التي يضعها اللغويون، وتستهلك منهم أوقات طويلة من العمل [5] مثل وجوب إتباع حروف الجر بالأسماء، ووجوب إتيان الفعل المضارع بعد أداة النفي "لم"، وتميُّز الفعل الماضي عن المضارع والأمر باتصاله بتاء الفاعل وتاء التأنيث الساكنة، ولكن مع كل هذه القواعد الضابطة لوسم الكلمة الصرفي فإنه يوجد سياقات أخرى لا تستطيع القواعد الحكم فيها بدقة ولا تستطيع ترجيح قسم من أقسام الكلام على آخر، فقواعد اللغة المعيارية لها مدى تعمل فيه بدقة لكنها لا يمكن الاعتماد عليها بمفردها في ضبط الوسم الصرفي لكل كلمة في النص.

ب- الاتجاه الوصفي المعتمد على المدونات وأنظمة تعليم الآلية

ويندرج تحت هذا الاتجاه النظام الموجَّه في تعليم الآلة، والنظام غير الموجَّه، والنظام شبه الموجّه. وبحلول عام 2000 وصلت الأنظمة الآلية الموجّهة المعتمِدة على تقنيات تعليم الآلة إلى درجة عالية من الدقة في معالجة الدلالة فضلًا عن معالجة الصرف وأقسام الكلام، واتجهت الأنظار نحو الأنظمة شبه الموجّهة، والأنظمة غير الموجّهة المعتمدة على المدونات اللغوية الصمّاء، والأنظمة المختلطة [6].

الأنظمة الموجهة في تعليم الآلة

تتصف الأنظمة الآلية الموجَّهة في تعليم الآلة بالاعتماد على عينة كبيرة من النصوص المحللة، والاستفادة منها في معالجة الإشكاليات اللغوية في نصوص جديدة غير محللة، ويكون ذلك من خلال الإحصائيات المستنبطة من العينة المحللة، أو من خلال تطويع العينة المحللة لتقنيات تعليم الآلة (MACHINE-LEARNING TECHNIQUES) بالتدرب عليها (TRAINING STAGE) من خلال نماذج رياضية وإحصائية، ثم اختبار نصوص جديدة غير محللة (TESTING STAGE).

وبشكل عام، حقق هذا الاتجاه في المعالجة نتائج أفضل من نظيره غير الموجّه [7]. ومن أشهر النماذج الإحصائية والخوارزميات التي تندرج تحت هذا الاتجاه طريقة مصنف BAYES البسيط، وطريقة آلة الدعم الموجّهة SVM، وطريقة نموذج MARKOV، وغيرها.

- **طريقة مصنف BAYES البسيط (NAIVE BAYES CLASSIFIER)**

هو حساب بسيط للاحتمال الشرْطي (CONDITIONAL PROBABILITY) لكل لقب صرفي من ألقاب الكلمة الملتبسة في ظل خصائص السياق المحيط، وتتلخص هذه الطريقة في تمثيل السياق الذي يحدث فيه الكلمة الملتبسة في شكل قوة موجّهة (VECTOR) من الخصائص (FEATURE VARIABLES) F=(f1, f2, . . . , fN)، وتمثيل الألقاب الصرفية المختلفة للكلمة في شكل مجموعة من المتغيرات (CLASSIFICATION VARIABLES) POS1, POS2, . . .,POSK)، فتكون عملية اختيار اللقب الصرفي المقصود هي اختيار للقب الذي يحقق أعلى احتمالية للحدوث في ظل وجود مجموعة خصائص السياق [5] [8]. وتعتمد تلك الطريقة على أن الكلمات المحيطة بالكلمة الملتبسة في السياق تساهم في تحديد وسمها الصرفي.

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

- **خوارزمية (KNN) THE K-NEAREST NEIGHBOR**

تستند هذه الطريقة في معالجة الالتباس على أمثلة محللة مسبقًا يتم مقارنة خصائصها بخصائص الجملة المختبَرة التي تحتوي على الكلمة المراد وسمها صرفيًا، ثم تحديد الأمثلة الأقرب في الخصائص للجملة المختبَرة (K-CLOSEST TRAINING EXAMPLES)، ويكون الوسم الصرفي المقصود للكلمة في الجملة المختبَرة هو الوسم الأكثر تكرارًا في الأمثلة الأقرب في الخصائص (K-EXAMPLES) [9].

- **التلقيب الصرفي باستخدام آلة المتجهات الداعمة (SVM) SUPPORT VECTOR MACHINE**

تُعد آلة المتجهات الداعمة مثالًا مباشرًا على أدوات تعليم الآلة الموجّهة التي تستخدم في التصنيف الآلي، فهي أحد المصنفات الثنائية (BINARY CLASSIFIER) المزودة بخوارزمية تعليم (LEARNING ALGORITHM) تقوم بالتدرّب على العينة المحللة وبناء نموذج يستطيع تحليل عينات جديدة غير محللة. وكان VAPNIK [10] أول من قدّم واستخدم هذا المصنّف، وقد طبقت تلك الطريقة في كثير من تطبيقات المعالجة الآلية للغة الطبيعية بشكل عام، وأثبتت نجاحات ملحوظة خاصة في تطبيقات تصنيف النصوص (TEXT CATEGORIZATION)، وتقسيم الجمل إلي مركبات (PHRASE CHUNKING) [11]، والإعراب الآلي (PARSING)، و فك التباس المعنى (WSD).

ويوصف SVM بأنه مصنّف ثنائي، فهو يعالج المشكلات ذات التصنيفات الثنائية (BINARY CLASSIfiCATION PROBLEMS) مما يستدعي تهيئته لمعالجة المشكلات المتعددة التصنيف (MULTI-CLASS CLASSIfiCATION PROBLEMS) مثل مشكلة الالتباس الصرفي، فقد يتواجد ألقاب صرفية متعددة للكلمة الواحدة [12].

الأنظمة شبه الموجَّهة في تعليم الآلة (RECURSIVE OPTIMIZATION)

وتتميز هذه النوعية من الأنظمة باستفادتها من أقل قدر من النصوص المحللة (SEED) واستخدامها في تحليل أمثلة جديدة غير محللة (وبالتالي أقل قدر من التدخل البشري بدلًا من الكميات الضخمة التي تتطلبها الطرق الموجّهة)، ثم إضافتها (بعد مراجعتها) إلى النصوص التدريبية المحللة من أجل تحسين (OPTIMIZING) أداء المحلل بزيادة النصوص التدريبية المحللة، ثم إعادة تدريب المصنّف، واستخدامه في تحليل نصوص جديدة، وهكذا بشكل متكرر (RECURSIVE) من أجل التحول تدريجيًا إلى الميكنة الآلية الكاملة، وهذا الأسلوب يستخدم في معالجة الالتباس الدلالي بشكل خاص، وبناء تطبيقات المعالجة الآلية بشكل عام. والشكل التالي يعرض آلية حدوث التحسين المتكرر لعينة صغيرة من النصوص المحللة.

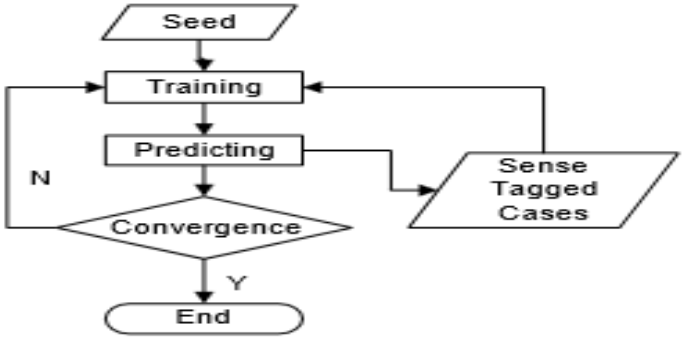


Figure 2. Flow of Recursive Optimization Algorithm

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

شكل 2. آلية حدوث التحسين المتكرر لعينة صغيرة من النصوص المحللة

الأنظمة غير الموجَّهة في تعليم الآلة

وتسمى أيضًا (DATA-DRIVEN - LANGUAGE-INDEPENDENT – KNOWLEDGE-LEAN APPROACHES)، وهذه النوعية من الطرق في معالجة الالتباس لا تعتمد على أي مصدر من مصادر المعلومات اللغوية أو المدونات المحللة، وإنما تعتمد فقط على المدونات الصمّاء (RAW CORPORA) الخالية من أي نوع من التحليل اللغوي، لذلك فهي توصف بأنها طرق هزيلة المعلومات (KNOWLEDGE-LEAN METHODS)، وهذا ما يميّزها عن النوعين السابقين، وبسبب ذلك فإنها تفتقر إلى التحديد الدقيق للقب الصرفي (ASSIGNING POS TAGS)، لافتقارها إلى المعلومات اللغوية الصرفية، فهي فقط تستطيع تمييز (DISCRIMINATION) الألقاب الصرفية المختلفة في فصائل يُطلق عليها اسم عناقيد (CLUSTERS) [5].

1. مدونة التدريب والاختبار

تعتمد أنظمة تعلّم الآلة الموجهة، ومنها SVM، بشكل أساسي على وجود مدونة محللة على المستوى اللغوي المراد اختباره، وهو المستوى الصرفي في هذا البحث، من أجل تدريب النظام عليه بشكل كافٍ يسمح له بالتنبؤ بشكل صحيح بالوسم الصرفي لكلمات جديدة في مرحلة الاختبار. وقد وقع الاختيار على جزء من المدونة العربية العالمية (ICA) [13] المحللة صرفيًا على مستوى أقسام الكلام وغيرها من المعلومات الصرفية والنحوية مثل النوع (gender)، والعدد (number)، والشخص (person)، والتعريف والتنكير، والزمن، والبناء للمعلوم والمجهول، والمدخل المعجمي، وغيرها من الخصائص. والمدونة العربية العالمية هي مدونة ممثلة للغة العربية المعاصرة المكتوبة، وقد تم تجميع مادة هذه المدونة من مصادر شتى للمحتوى العربي بمختلف مجالاته وأقطاره لتكون بذلك محاكية للمدونة الإنجليزية العالمية (International Corpus of English) من أجل دعم البحث اللغوي، وقد خُطّط لهذه المدونة أن تبلغ مائة مليون كلمة. وقد انفردت المدونة العربية العالمية بمجموعة من الخصائص التي جعلتها في مقدمة المدونات العربية الأوسع انتشارًا والأكثر استخدامًا، هذه الخصائص هي:

- إتاحتها للباحثين مجانًا من خلال موقع مكتبة الإسكندرية (http://bibalex.org/ica/ar/).
- أنها محللة على المستوى الصرفي مما يمكّن الباحث من تحديد البحث داخلها وبالتالي سهولة الحصول على نتائج البحث المرادة.
- أنها ممثلة للغة العربية المعاصرة ومتوازنة من حيث حجم المجالات المكونة لها ومطابقة ذلك للواقع اللغوي، وبالتالي يمكن تعميم نتائج الأبحاث المبنية عليها على اللغة المعاصرة.
- مراعاة أن تكون هذه المدونة متعددة الأغراض وليست مخصصة لنوعية محددة من الدراسات، وقد روعي ذلك أثناء التخطيط لبنائها وتجميع مادتها لتكون صالحة لجميع الدراسات والأبحاث اللغوية المختلفة.

وقد وقع الاختيار على عينة لغوية حجمها مائة ألف كلمة تقريبًا من أجل تقسيمها إلى جزء للتدريب، وجزء للاختبار.

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

2. قائمة أقسام الكلام (POS)

بلغ عدد الألقاب الصرفية المستخدمة في التجربة 48 لقبًا صرفيًا، وهي نفس قائمة أقسام الكلام المستخدمة في تحليل المدونة العربية العالمية، وهي مستنبطة من أقسام الكلام المستخدمة في بناء المحلل الصرفي العربي TIM BUCKWALTER مع بعض التعديلات والإضافات. وتنقسم قائمة الألقاب الصرفية إلى مجموعة الألقاب اللغوية المشتملة على الأسماء وأقسامها الفرعية، والأفعال وأقسامها الفرعية، والادوات، والروابط، والحروف وأقسامها الدلالية، والكلمات العامية، والدخيلة، والأجنبية، والاختصارات، إلى جانب مجموعة الألقاب غير اللغوية التي تشمل علامات الترقيم، والأرقام، والتواريخ، والروابط الاليكترونية، والعلامات، والرموز، والعناوين الإليكتروني، وتشمل أيضًا مجموعة الأكواد الترميزية الهيكلية للنص (STRUCTURAL MARK UP CODES) مثل بداية النص ونهايته، وبداية الفقرة ونهايتها، وبدايه العنوان الرئيسي والفرعي ونهايته. والجدول التالي يضم مجموعة أقسام الكلام اللغوية وغير اللغوية المستخدمة في البحث.

جدول 1. قائمة أقسام الكلام (POS) المستخدمة في التلقيب الآلي – وعددها 48 لقبًا صرفيًا

ADJ	NOUN _PROP (ADV_T)	NOUN_ PROP	NOUN (VERBAL)	NOUN (ADV_T)	NOUN (ADV_P)	NOUN (ADV_M)	NOUN
			REL_PRON	DEM_PRO N	PRON_3MS	PRON_1S	PRON
			PV_PASS	PV	IV_PASS	IV	CV
INTERJ	VERB_PAR T	NEG _PART	INTERROG_P ART	FUT _PART	FOCUS _PART	EXCEPT _PART	PART
		ABBRE V	Not _Arabic	Colloquial	SUB_CONJ	CONJ	PREP
Exl_Mark	Qus_Mark	Email	Date	Site	Sign	Punc	Num
				BOF_Tit	BOF_Qus	BOF_Prg	BOF_Doc
				EOF_Tit	EOF_Qus	EOF_Prg	EOF_Doc

أداة التعلّم الآلي YAMCHA

هي أداة متعددة المهام ومفتوحة المصدر تقوم بأداء الكثير من مهام معالجة اللغة، مثل التلقيب الصرفي الآلي للكلمات، والتعرف على أسماء الكيانات، والتحليل التركيبي للجمل، وغيرها من المهام اللغوية. وتعتمد هذه الأداة في التعلم الآلي على آلة المتجهات الداعمة (SVM) التي تستخدم في تصنيف البيانات بدقة وكفاءة بالغة لأنها تستخدم جزء من البيانات في التدريب والتعلم [14]. ولكي تتم عملية التدريب بشكل صحيح يجب أن يتم إعادة صياغة عينة التدريب في شكل أعمدة معروضة في ملف نصي بسيط، كما هو موضح في شكل (3).

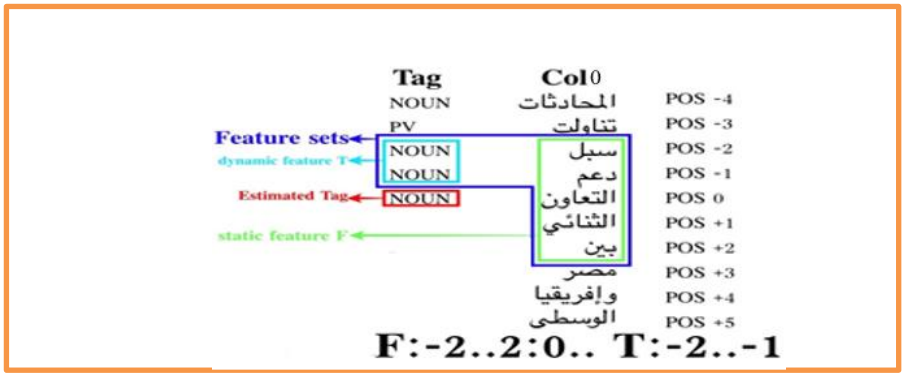
INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

شكل3. صياغة عينة التدريب في شكل أعمدة معروضة في ملف نصي بسيط

ويتيح نظام YAMCHA تغيير مدى ونوع المعلومات اللغوية المعتمد عليها في التعلم الآلي (feature set and window-size) من خلال معامل الخصائص (features parameter) في أمر التدريب (training command). فأمر التدريب التالي :

```
make CORPUS=train.data MODEL=case_study FEATURE="F:-2..2:0.. T:-2..-1"
train
```

يوضح الأخذ في الاعتبار الكلمات التي في المدى [-2: 2] (أي كلمتان قبلُ وكلمتان بعدَ الكلمة المراد تحليلها) أثناء التدريب، وكذلك الألقاب الصرفية في المدى [-2 : -1] (أي لقبان صرفيان قبل الكلمة المراد تحليلها) كما هو موضح في الشكل التالي:



شكل 4. تدريب المصنّف على عينة لغوية مع الأخذ في الاعتبار الكلمات التي في المدى [2- : 2]، والألقاب الصرفية في المدى [2- : 1-
[

مرحلة تدريب المدونة

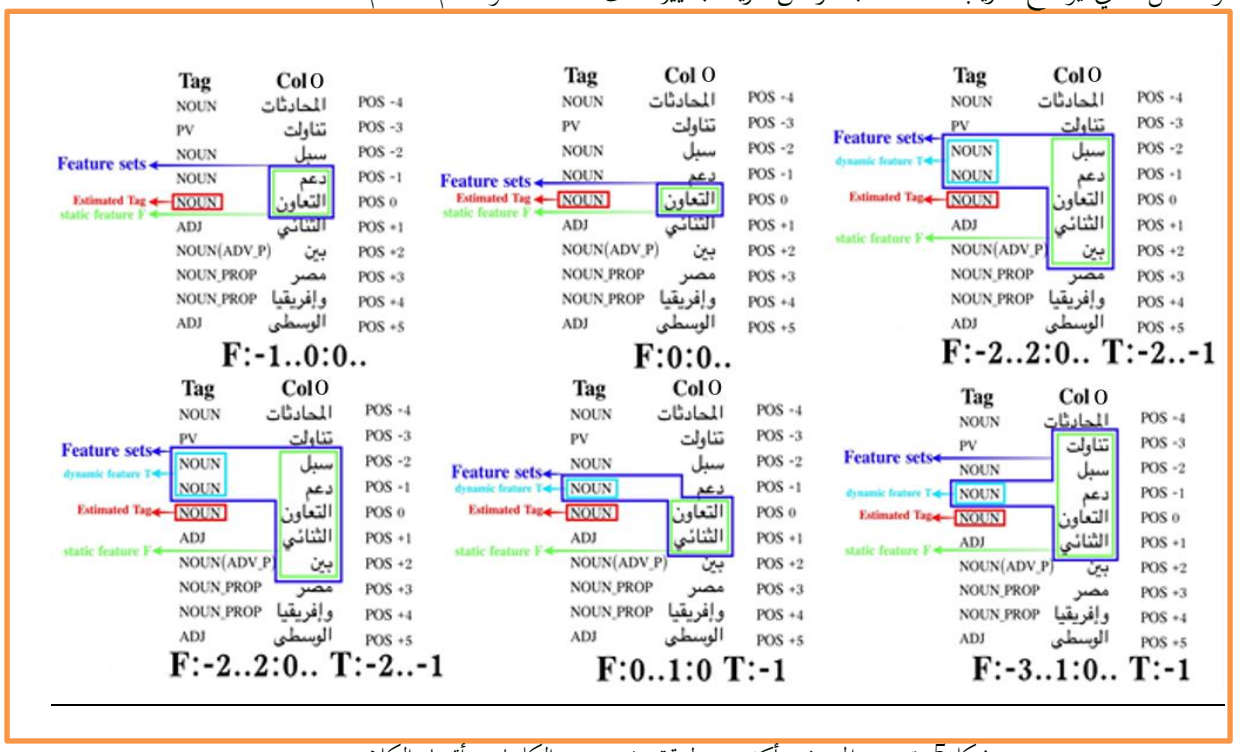
تمّ تقسيم مدونة البحث، وحجمها 100,039 كلمة، إلى جزء خاص بتدريب مصنف SVM الثنائي، وقدره 64.608كلمة (70% من حجم المدونة)، وجزء خاص باختبار نصوص جديدة وتقييم النتائج، وقدره 35.431 كلمة

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

(30% من حجم المدونة). وقد تمّ تدريب المصنِّف عدة مرات، وفي كل مرة يتمّ تغيير مدى التدريب الخاص بالكلمات وأقسام الكلام من خلال أمر التدريب السابق لتتولد ملفات التدريب التالية:

case_study.log	:	log of training
case_study.model	:	model file (binary, architecture dependent)
case_study.txtmodel.gz	:	model file (text, architecture independent)
case_study.se	:	support examples
case_study.svmdata	:	training data for SVMs

والشكل التالي يوضح تدريب المصنف بأكثر من طريقة بتغيير مدى الكلمات وأقسام الكلام

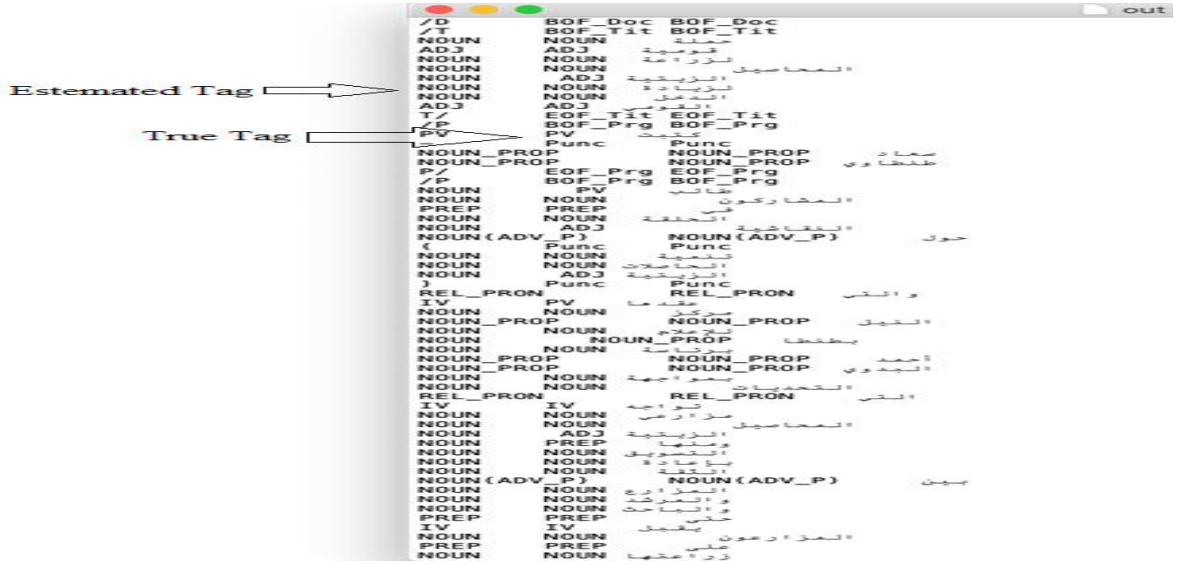


شكل5. تدريب المصنف بأكثر من طريقة بتغيير مدى الكلمات وأقسام الكلام

مرحلة اختبار نصوص جديدة

في هذه المرحلة تمّ اختبار 30% من مدونة البحث بالاعتماد على التدريب السابق على جزء من المدونة (70%)، وتمّ إعادة صياغة شكل مدونة الاختبار بنفس شكل مدونة التدريب (في شكل أعمدة متوازية من الكلمات والوسم الصرفي معروضة في ملف نصي بسيط) من إجل إتمام عملية التلقيب الآلي، ثم تمّ تقييم عملية التلقيب الآلي بمقارنة التلقيب الصحيح للكلمات بالتلقيب المقترح من قِبَل المصنف الآلي، كما هو موضح في شكل (6).

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023



شكل 6. مرحلة اختبار نصوص جديدة وتقييم عملية الاختبار

النتائج

بعد إجراء عملية التدريب عدة مرات مع تغيير مدى الكلمات وأقسام الكلام ثم اختبار نصوص جديدة وتقييمها، تم رصد نسبة خطأ التلقيب الآلي في كل تجربة كما هو موضح في جدول (2) التالي والرسم التوضيحي (7) الذي يليه:

جدول 2. رصد نسبة خطأ التلقيب الآلي في كل تجربة مع تغيير مدى المعلومات المعتمد عليها في التدريب

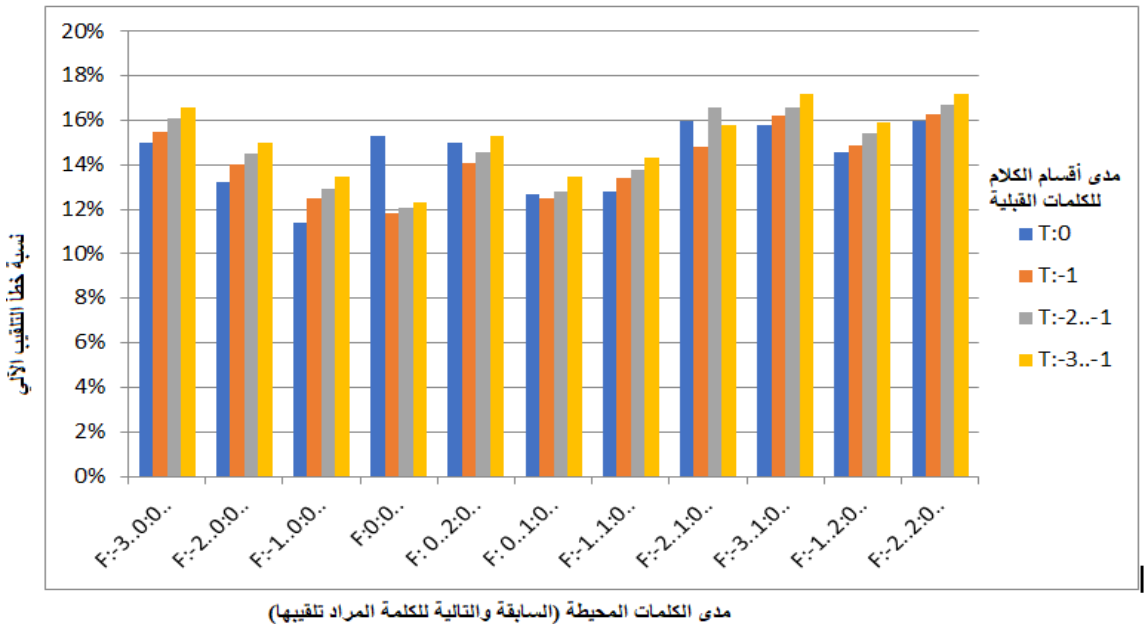
F \ T	T:0	T:-1	T:-2..-1	T:-3..-1
F:-3..0:0..	15%	15.5%	16.1%	16.6%
F:-2..0:0..	13.2%	14%	14.5%	15%
F:-1..0:0..	11.4%	12.5%	12.9%	13.5%
F:0:0..	15.3%	11.8%	12.1%	12.3%
F: 0..2:0..	15%	14.1%	14.6%	15.3%
F: 0..1:0..	12.7%	12.5%	12.8%	13.5%
F:-1..1:0..	12.8%	13.4%	13.8%	14.3%
F:-2..1:0..	16%	14.8%	16.6%	15.8%
F:-3..1:0..	15.8%	16.2%	16.6%	17.2%
F:-1..2:0..	14.6%	14.9%	15.4%	15.9%
F:-2..2:0..	16%	16.3%	16.7%	17.2%

ونلاحظ من النتائج السابقة أن :

- أقل نسبة خطأ تم رصدها هي (11.4%)، وكانت في حالة اعتبار الكلمة السابقة في التحليل دون النظر إلى لقبها الصرفي (F:-1..0:0..).
- كلما زاد مدى الكلمات البَعدي (الكلمات التالية للكلمة المراد تحليلها) أو القبلي، كلما زادت نسبة الخطأ.

INSANIAH: Online Journal of Language, Communication, and Humanities
Volume 6 (1), April 2023

- الكلمة السابقة للكلمة المراد تحليها أكثر إفادة ودلالة على اللقب الصرفي للكلمة المراد تحليلها من الكلمة التالية.
- كلما زاد مدى الألقاب الصرفية القبلية، كلما زادت نسبة الخطأ ولكن ليس بنفس زيادة نسبة الخطأ في حالة النظر إلى الكلمات المحيطة القبلية والبعدية.
- بشكلٍ عام، زيادة مدى المعلومات اللغوية المعتمد عليها في التدريب يؤدي إلى زيادة نسبة الخطأ.
- الألقاب الصرفية للكلمات السابقة للكلمة المراد تحليلها أكثر إفادة من الكلمات المحيطة بها.
- ترتيب المعلومات اللغوية المساعدة في التنبؤ باللقب الصرفي حسب أهميتها هو: الكلمة السابقة (F:-1..0:0..T:0) ثم اللقب الصرفي للكلمة السابقة (F:0:0..T:-1) ثم اللقبان الصرفيان للكلمتين السابقتين (F:0:0..T:-2..-1).



شكل 7. يوضح نسبة خطأ التلقيب الآلي في التجارب المختلفة مع تغيير مدى المعلومات اللغوية المعتمد عليها في التلقيب

المصادر والمراجع

Alansary, Sameh, and Magdi Nagi. "The international corpus of Arabic: Compilation, analysis and evaluation." Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). 2014

Alhasan, Ahmad, and Ahmad T. Al-Taani. "POS tagging for arabic text using bee colony algorithm." Procedia computer science 142 (2018): 158-165.

Crystal, David. An encyclopedic dictionary of language and languages. Penguin, 1994.

Edmonds, Philip and Agirre, Eneko. "Word sense disambiguation". Scholarpedia 3(7):4358 (2008).

Garside, Roger. "The CLAWS word-tagging system." The Computational analysis of English: A corpus-based approach. London: Longman (1987): 30-41

Greene, Barbara B., and Gerald M. Rubin. Automatic grammatical tagging of English. Department of Linguistics, Brown University, 1971

Habash, Nizar, and Owen Rambow. "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop." Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05). 2005.

Johnson, Mark. "How relevant is linguistics to computational linguistics." *Linguistic Issues in Language Technology* 6.7 (2011): 1-23.

Kudo, Taku, and Yuji Matsumoto. "Fast methods for kernel-based text analysis." Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003.

Kudo, Taku, and Yuji Matsumoto. "Chunking with support vector machines." Second Meeting of the North American Chapter of the Association for Computational Linguistics. 2001.

Le, Cuong Anh, and Akira Shimazu. "High WSD accuracy using Naive Bayesian classifier with rich features." Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation. 2004.

Mahafdah, R. Nazlia O. and Al-Omari, O. (2014) Arabic part of speech tagging using K-Nearest Neighbor and Naive Bayes classifiers combination Journal of Computer Science, 10(10):1865-1873.

Mitkov, Ruslan, ed. The Oxford handbook of computational linguistics. chapter 24, corpus linguistics, by Tony Mcenery. Oxford University Press, 2004.

Navigli, Roberto. "Word sense disambiguation: A survey." ACM computing surveys (CSUR) 41.2 (2009): 1-69.

Ng, Hwee Tou, and Hian Beng Lee. "Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach." arXiv preprint cmp-lg/9606032 (1996).

Vladimir N. Vapnik. 1995.The Nature of Statistical LearningTheory. Springer.

Yousif, Jabar, and Maryam Al-Risi. "Part of Speech Tagger for Arabic Text Based Support Vector Machines: A Review." ICTACT Journal on Soft Computing: DOI 10 (2019).