العرب
الإثنين 2019/09/23
السنة 42 العدد 11476

تطوير تطبيقات للغة العربية يواجه تحديات تقنية كثيرة

برنامج التعرف التلقائي على اللهجات يهدف إلى تطوير موارد وتقنيات للهجات 25 مدينة عربية

تطرح مسألة معالجة اللغات الحية عن طريق الخوارزميات الكثير من التحديات يعود سببها الأساسي إلى خصوصية البعض من اللهجات لا سيما في اللغات التي تتوفر بشأنها البيانات اللازمة ضمن أنظمة الذكاء الاصطناعي المختلفة ومن بين هذه اللغات العربية التي يواجه الباحثون العديد من التحديات على طريق تطوير تطبيقات بشأنها.

محمد الحمامصي

تعد معالجة اللغات الطبيعية مجالا واسعا متعدد التخصصات يربط بين علم الكمبيوتر واللسانيات والرياضيات وعلم النفس. ولهذا المجال أهداف بحثية عديدة نظرية وتطبيقية من بينها نمذجة كيفية اكتساب الإنسان للغة إلى برمجة تطبيقات ذكاء اصطناعي كالترجمة الآلية والتعرف الآلي على الكلام والمساعدات الآلية أو الشات بوتات مثل "سيري" و"ألكسا".

وجزء كبير من هذا المجال يعمل على ما يسمى بالتقنيات التمكينية كالتحليل الصرفي الآلي مثلا الذي يُستخدم كعنصر في بناء التطبيقات الذكية.

أجرت "العرب" حوارا مع البروفيسور نزار حبش وهو من أبرز المتخصصين في معالجة اللغات الطبيعية واللسانيات الحاسوبية وتشمل اهتماماته البحثية الترجمة الآلية والتحليل الصرفي والتحليل النحوي والنمذجة الحاسوبية للغة العربية ولهجاتها المرتبطة بها.

يعتبر حبش، مدير مختبر المقاربات الحاسوبية لنمذجة اللغة في جامعة نيويورك أبوظبي، أن معالجة اللغات الطبيعية، أو ما يسمى أحيانا باللسانيات الحاسوبية أو تقنيات اللغة، فرع من حقل الذكاء الاصطناعي يركز على تحليل وتوليد ونمذجة اللغات البشرية الطبيعية كالعربية والفرنسية بعكس اللغات المصطنعة كلغات البرمجة مثلا.

موارد البيانات

لكون معالجة اللغات الطبيعية فرعا من الذكاء الاصطناعي، يوضح حبش أن الأساليب الأكثر استخداما فيه اليوم هي أساليب التعلم الآلي العميق والإحصائي والتي تعتمد بشدة على موارد البيانات. فمثلا، لبناء نظام ترجمة آلي بين لغتين يلزمنا الآلاف بل حتى الملايين من الجمل المترجمة في اللغتين "أو ما يسمى بالمدونات المتوازية" والتي تبني خوارزميات التعلم الآلي منها نماذج حاسوبية للترجمة.

وبالطبع، فإن جودة النتائج تعتمد على نوعية الخوارزميات وكمية البيانات المتوفرة لها. وعلى الرغم من كون الخوارزميات والأساليب المستخدمة في معالجة اللغات الطبيعية حيادية لغويا، إلا أننا نجد أن جودة التطبيقات أعلى في لغات معينة كالإنكليزية والألمانية لكونها غنية بموارد البيانات وليس لأي خصائص لغوية تميزها عن لغات فقيرة نسبيا بموارد البيانات كاللغة العربية ولهجاتها.

وأشار حبش إلى أنه باعتباره باحثا عربيا في الذكاء الاصطناعي اللغوي فإنه يسعى لتوصيف الوضع الحالي للعربية بموضوعية لا انتقادية. وقال "اللغة العربية مجموعة من اللهجات المرتبطة ببعضها لغويا وتاريخيا، والتي يتميز بعضها اجتماعيا.

لغة الإنسان العربي اليومية قد يتخللها تنقل بين العامية والفصحى ومفردات على درجات مختلفة من التعرّب من لغات أخرى كالإنكليزية والفرنسية. ويرى حبش أن تطبيقات الذكاء الاصطناعي اللغوية يجب أن تُصمم لخدمة هذا الإنسان واحتياجاته اللغوية كلها: مثلا، أنظمة المساعدة الآلية يجب أن تتعرف تلقائيا على لهجته، وتتفهم أوامره وأسئلته، وتجيب عليها باللهجة التي هو يفضلها كمستخدم. واعتبر حبش أنه "من ناحية تكنولوجية، قد تكون اليوم خيالا علميا".

وقال "يجب علينا أن نبقي توقعاتنا عالية برغم التحديات التقنية التي تواجهنا في تطوير تطبيقات ذكية للغة العربية، وهي تحديات عديدة من بينها الصرف الغني جدا، الإملاء المبهم بسبب التشكيل الاختياري، قلة المدونات وموارد البيانات، تعدد اللهجات العامية، الفرق الكبير بينها وبين الفصحى، عدم وجود إملاء قياسي للعامية، وكثرة الأخطاء الإملائية في الفصحى كما نجدها اليوم على مواقع التواصل الاجتماعي".

ويرى أن هذه التحديات مفردة ليست خاصة باللغة العربية، ولكن تجمُّعها معا يجعل العمل مع اللغة العربية صعبا نسبيا.

ويهدف برنامج حبش البحثي والتطبيقي طويل المدى والمركز على اللغة العربية إلى إيجاد حلول لهذه الصعوبات وبناء تطبيقات وتقنيات تمكينية وكذلك موارد بيانات متخصصة لتسهيل العمل على تطبيقات الذكاء الاصطناعي للعربية.

ويؤكد حبش أن اللغة العربية لغة مفعمة بالحياة، وأن اللغات الحية هي في حالة تغيير وتطور مستمرين، بعكس اللغات التي لا تتغير وتتماشى مع العصر والتي عادة تكون لغات بائدة أو مصطنعة.

ولكن من ناحية مكانة العربية ونفوذها المرتبطين بقوة متكلميها السياسية، يقول "يمكنني قبول فكرة تدهورها نسبيا مقارنة مع مكانتها العالمية في العصور الوسطى، ولكنني لا أرى تدهورا مستمرا يوما بعد الآخر خلال المئة عام الأخيرة".

ويوضح أنه بالعكس فنسبة الأمية في العالم العربي ككل قد انخفضت مما كانت عليه وخصوصا للمرأة العربية، وإن مازالت الأمية مشكلة مقلقة في بعض البلدان العربية. وترتبط مكانة اللغة كذلك بدرجة التقدير الذي يكنه لها متكلموها واهتمامهم بها وبتعليمها.

ومن هذه الناحية، يرى حبش أن لدى أبناء العربية حالة انفصام لغوي وعلاقة حب وكره مع لغتهم، فتراهم يقدسون الفصحى ثم يهزأون ممن يحاول تكلمها في الحياة اليومية، وهم كذلك يتكلمون ويفكرون بالعامية ولكنهم يعتبرونها خاطئة مبتذلة وخالية من القواعد.

وهذا الوضع أسوأ أحيانا بين فئات في الطبقات العليا التي يتجنب بعضها التكلم بالعربية مفضلين الإنكليزية أو الفرنسية. ويؤكد حبش "أنا لست رجل سياسة ولا يمكنني أن أتكلم عن كيفية حل هذا الموضوع، ولكنني كمتخصص في اللسانيات والحوسبة أرى أن تحسين علاقتنا مع لغتنا وقبول تعدديتها والعمل على تحسين أنظمة تعليمها لأبنائها وللأجانب المهتمين بها وبناء تطبيقات ذكية لدعمها وخدمتها بكل أشكالها، فصحى وعامية، خطوات في الاتجاه السليم لرفع مكانتها".

> جودة التطبيقات أعلى في لغات معينة كالإنكليزية والألمانية لكونها غنية بموارد البيانات وليس لخصائص لغوية تميزها عن لغات فقيرة نسبيا بموارد البيانات كاللغة العربية ولهجاتها

ويعتبر حبش أن الإملاء والصرف هما أبرز تحديات اللغة العربية في مجال معالجة اللغات الطبيعية. أما بالنسبة للنحو، وبنية الجملة ودلالتها، فاللغة العربية ليست فريدة من نوعها بين لغات العالم. وهذا لا يعني أنها مسائل سهلة الحل، على العكس، بل وتحتاج إلى الكثير من الموارد الخاصة، ولكنها ليست فريدة ومميزة للغة العربية.

أولى تحديات الإملاء العربي هي التشكيل الاختياري المسبب للإبهام. فعدا النصوص المقدسة وكتب الأطفال المدرسية، نجد أن التشكيل المساعد لفك اللبس الدلالي والنحوي يستخدم بنسب ضئيلة جدا، لا تتعدى 1.5 بالمئة في نصوص الجرائد اليومية.

وطبعا التشكيل ليس ضروريا للشخص المتعلم العربي، ولكنه يسبب المشكلات لبرامج الذكاء الاصطناعي التي تستخدم السياق لفك الإبهام. فمثلا كلمة "وحده" لها أكثر من 200 تحليل موزعة على 12 مدخلا معجميا في نظام التحليل العربي "كلمة-ستار" المبني على نظام "ساما" للصرف العربي. من بين هذه التحاليل، نجد "وَ+ حَدُ +هُ"، "وَحْدَ+هُ"، "وَحَّدَ+هُ"، و"وَحْدَة" "المثال الأخير بالتاء المربوطة يعتبر الكلمة المدخلة غير صحيحة إملائيا".

وأوضح أن مشكلة الإملاء أكثر تعقيدا بالنسبة للهجات العربية لأنه لا توجد قواعد إملائية قياسية لها تُدرس في المدارس. ومع أن اللهجات كانت عادة توصف بأنها فقط محكية، إلا أنه في عصر وسائط التواصل الاجتماعي والهواتف الذكية، فقد أصبحت اللهجات العامية فعليا مكتوبة أكثر من الفصحى.

وعدم وجود إملاء قياسي مدعوم من الأسس التعليمية واللغوية يؤدي إلى تفشي الأشكال الإملائية في العامية. في أحد الأبحاث، وجد حبش وزملاؤه 27 تهجئة مختلفة لكلمة "مبيقولهاش" المصرية.

وهذا بدوره يضعف جودة أنظمة الذكاء الاصطناعي ويجعل تقييمها صعبا أو شبه مستحيل، حيث لا يوجد اتفاق على النتيجة الصحيحة المطلوبة لبرامج التعرف الآلي على الكلام، مثلا.

وأخيرا، في كل من العربية الفصحى والعاميات، هناك نظام صرفي معقد جدا تنتج عنه تراكيب صرفية كثيرة: على سبيل المثال الفعل "وسيكتبونها" هو واحد من حوالي 5.400 تصريف للفعل "كتب".

وفي حين تبسط اللهجات العامية الصرف من حيث إسقاط الحالة الإعرابية إلا أنها تضيف العديد من الأدوات المتصلة: على سبيل المثال في اللهجة المصرية، هناك قاعدة بسيطة لنفي أي فعل بإضافة "ما" و"ش": قال-ماقالش، شاف-ماشافش، قالهالي-ماقالهاليش، وهكذا.

هذه الدرجة العالية من التباين والغموض الإملائيين والثراء الصرفي، معا، تشكل حالة فريدة من نوعها إلى حد ما. في عمل متخصصي الذكاء الاصطناعي حول هذه المشكلة، احتاج هؤلاء إلى إنشاء مقاييس إملائية وموارد وخوارزميات.

أولا، من حيث المقاييس الإملائية، قام حبش وفريقه بتطوير نظام الكتابة الواحدة للدارجة "كودا". يحدد هذا النظام إملاء قياسيا لكتابة اللهجات العربية، ليمكن مجمع الباحثين من تطوير الموارد وتقييم الأنظمة على نحو متسق.

ثانيا، من حيث الموارد، شارك الباحثون وقاموا بتطوير عدد من موارد البيانات الموسّمة للهجات المصرية والفلسطينية والعراقية واليمنية والنجدية وأخيرا الخليجية بالتعاون مع عدد من المؤسسات الدولية.

وأخيرا عملوا على التحليل الصرفي الإملائي في نظام "مداميرا" وطوروه مؤخرا باستخدام تقنيات التعلم الشبكي العصبي العميق لخدمة المزيد من اللهجات وبجودة عالية.

أنظمة ذكية غير عنصرية

نبّه حبش إلى أن معظم أنظمة الذكاء الاصطناعي اللغوي تعتمد على موارد بيانات تعكس طبيعة المجتمعات البشرية بكل ما هو جيد وسيء فيها بما في ذلك التمييز الجنسي.

وهناك حاليا اهتمام كبير بين متخصصي الذكاء الاصطناعي بتطوير أنظمة ذكية غير عنصرية أو تمييزية. فمثلا عند ترجمة عبارتي "I am a doctor" و"I am a nurse" إلى العربية باستخدام مترجم غوغل نحصل على "أنا طبيب" و"أنا ممرضة"، ولا يمكننا الحصول على "أنا طبيبة" أو "أنا ممرض".

وفي بحث علمي تم تقديمه هذا الصيف عرض حبش وزملاؤه حلا يعتمد على تمييز جنس ضمير المتكلم وتوليد العبارة المترجمة بالضمير المذكر وكذلك المؤنث، وبهذه الطريقة يتوفر للمستخدم الخياران.

ووفق حبش، كجزء من هذا البحث ولكي يتمكن الباحثون من دراسة وتطوير أنظمة الترجمة وتقييمها، تم بناء أول مدونة متوازية للتأنيث والتذكير الصرفي "هي الأولى من نوعها ليس فقط للغة العربية".

تحتوي المدونة على 12 ألف جملة تم توسيمها كجمل يمكن استخدامها من قبل الضمير المتكلم المذكر أو المؤنث، أو كلاهما. وكانت نتائج البحث إيجابية حيث تمكننا من رفع جودة الترجمة الآلية، ولكن مازال هناك مجال كبير للتحسن. ونحن نعمل حاليا على توسيعها وتحسين نتائج استخدامها.

وكشف حبش أن "برنامج التعرف التلقائي على اللهجات العربية" الذي قام به فريق من الباحثين تحت إشرافه، أحد نتائج مشروع "مدار". والهدف منه هو تطوير موارد وتقنيات تمكينية وتطبيقات للهجات 25 مدينة عربية من الرباط إلى مسقط، وأيضا الفصحى.

وتتضمن الموارد التي قاموا بتطويرها حوالي 140 ألف جملة في نطاق السياحة مثلا، "كم تكلفة الإفطار؟"، بعضها متواز بخمس لهجات، وبعضها متواز بـ25 لهجة. كما قام الباحثون بوضع أول قاموس حاسوبي من 25 لهجة عربية لأكثر من ألف مفردة تقابلها مفردات الفصحى والإنكليزية والفرنسية.

وأخيرا، جمعوا عددا كبيرا من التغريدات العربية من ثلاثة آلاف مغرد تقريبا وتم وسمها بدولة المغرد. أما بالنسبة للتطبيقات، فهناك اثنان: الأول يركز على الترجمة الآلية، والآخر لتحديد أو تعريف لهجة النصوص المكتوبة تلقائيا. تتيح هذه التقنية للكمبيوتر التعرف على لهجة المستخدم بحيث يمكن استخدامها كميزة في النظام لتحسين الأداء.

ولفت حبش إلى أنه في صيف 2019 عُقد أكبر مؤتمرات اللسانيات الحاسوبية في فلورنسا، إيطاليا، وبلغ عدد الحضور حوالي 3.700 شخص كان من بينهم 27 فقط من دول عربية.

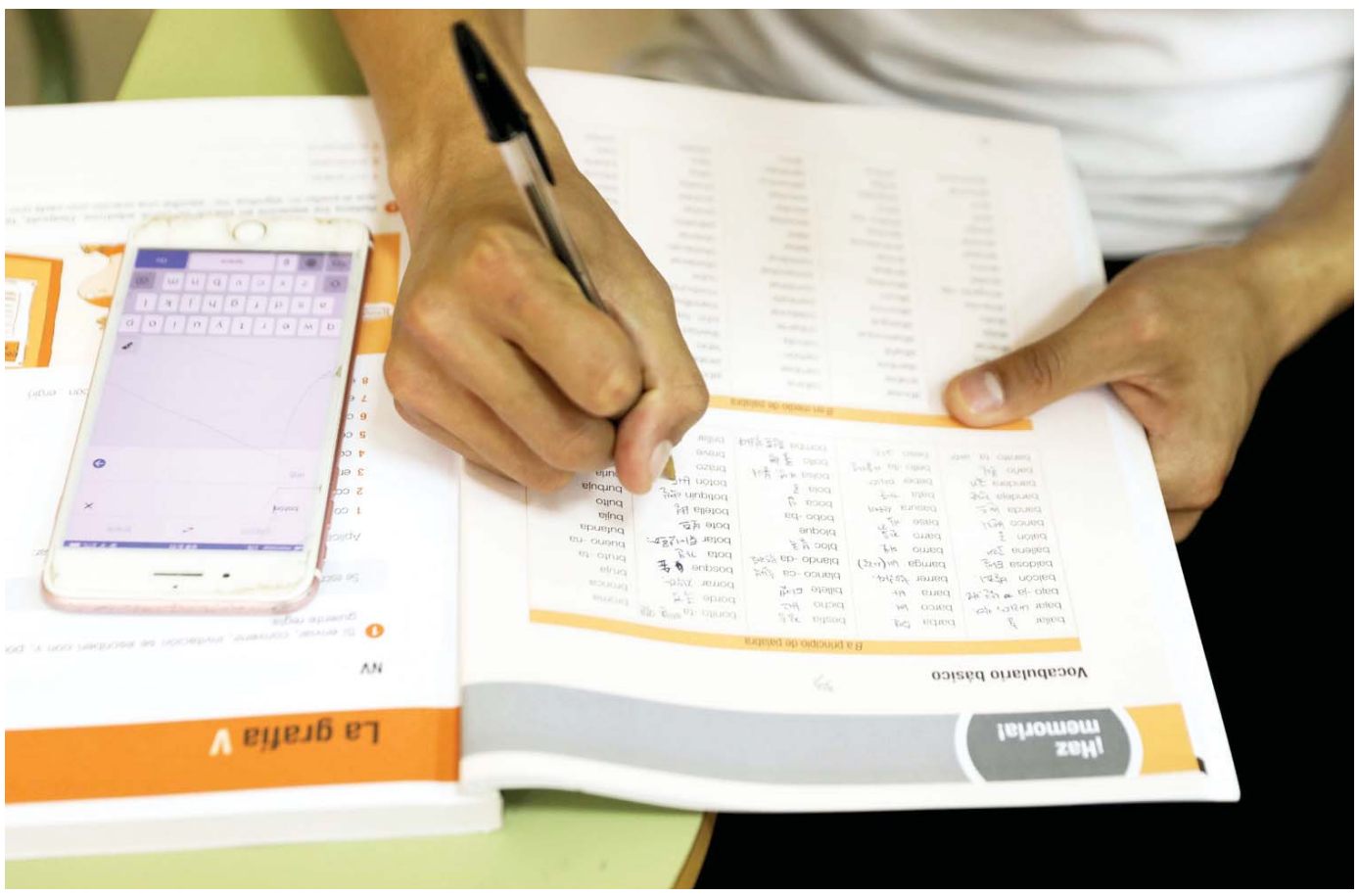
يتوقع حبش أن هذا الغياب سببه قلة دعم البحث العلمي في مجال الذكاء الاصطناعي في العالم العربي. وكمثال آخر، حتى عام 2018 كانت هناك أربعة موارد بيانات موسمة نحويا للغة العربية وجميعها تم تطويره خارج العالم العربي: اثنان في الولايات المتحدة، كان حبش المشرف على أحدهما، وواحد في المملكة المتحدة وواحد في جمهورية التشيك. هذه الموارد مهمة جدا للبحث العلمي ولتطوير تطبيقات الذكاء الاصطناعي اللغوي.

تطوير مورد بيانات عربي

في عام 2018 بدأت جامعة نيويورك أبوظبي بإنجاز مشروع "أرابك بالم تريبانك" الذي يهدف إلى تطوير مورد بيانات نحوي عربي يشمل نطاقا واسعا من النصوص العربية من المعلقات إلى تويتر.

وفي العالم العربي هناك عدد غير قليل من الباحثين الأكاديميين المهتمين بتقنيات اللغات ولكنهم ليسوا مدعومين كفاية ماديا وزمنيا. يقول حبش "بهذا أعني أن توفر لهم جامعاتهم زمنا كافيا ودعما ماليا لإجراء البحوث مقارنة مع مهام التدريس". أما بالنسبة لمطوري البرامج في الشركات في العالم العربي فهم على الأغلب لا ينشرون نتائجهم العلمية ولا يوفرون أدواتهم ومواردهم للآخرين الذين يعتبرون منافسين لهم.

شدد حبش على أن تخصص الترجمة الآلية يعتبر من أكثر التخصصات تقدما في مجال المعالجة الطبيعية للغات. حاليا، أساليب التعلم العصبي الشبكي العميق وكذلك التعلم الآلي الإحصائي هي الأكثر استخداما.



نظام ذكي لترجمة فورية

نزار حبش

> الإملاء والصرف أبرز تحديات اللغة العربية في مجال معالجة اللغات الطبيعية عن طريق الذكاء الاصطناعي

وبالنسبة للغة العربية هناك كمية كافية من هذه الموارد بالفصحى في نطاق الأخبار أو القرارات السياسية لتحقيق جودة عالية، ولكن كمية الموارد قليلة في نطاقات أخرى كاللغة الأدبية أو في اللهجات العامية.

لهذا، أنظمة الترجمة الآلية من غوغل أو فيسبوك تنتج جودة عالية في ترجمة الأخبار وجودة متدنية في ترجمة التعليقات والتغريدات عامة والعامية منها خاصة.

وهناك العديد من الأبحاث التي تبين أن التحليل الصرفي التقطيعي يساعد في رفع جودة الترجمة خاصة في حالات فقر الموارد. الفكرة هنا أن نظام الترجمة عندما يصادف كلمة معقدة صرفيا مثلا "ولبيتيه" (و+ ل+ بيتي +ه) فقد لا يتعرف عليها، ولكن عند تقطيعها كخطوة آلية سابقة للإدخال، فإننا نمكن النظام من ترجمتها.

كجزء من مشروع "مدار" تم جمع 20 ألف جملة في نطاق السياحة في خمس لهجات عربية (مغربية، تونسية، مصرية، شامية وخليجية) متوازية مع مدونة سبق بناؤها في الفصحى، الإنكليزية والفرنسية وتم نشر نتائج ترجمة عليها مؤخرا.

كذلك تمكن الباحثون من بناء مورد متواز لجزء من قرارات البرلمان الأوروبي مع اللغة العربية الفصحى وقيّم الترجمة الآلية بين 22 لغة أوروبية والفصحى بالاتجاهين للمرة الأولى.