اللهجات العربية والذكاء الاصطناعي: تحديات تكنولوجيا الصوت في بلادنا

ترجمة وتحرير: نون بوست

لا تستطيع أليكسا ولا كورتانا التحدث باللغة العربية، كما أن سيري لا تستطيع سوى فهم اللغة العربية الفصحى، بالإضافة إلى أن ترجمة غوغل بالكاد تكون دقيقة. وعندما يتعلق الأمر بفهم خامس أكثر لغة متحدّثٌ بها في العالم، فإن تكنولوجيا القرن الواحد والعشرين متأخّرة في هذا المجال. ومن جهته، قال عالم الكمبيوتر في جامعة بيرزيت في رام االله، مصطفى جرار: "يتحدث حوالي 300 مليون شخص في جميع أنحاء العالم اللغة العربيّة، وهي اللغة التي يستعملها قرابة 1.5 مليار شخص في دينهم، لكنها إحدى اللغات الأقل استخداما في التكنولوجيا".

شهد العقد الماضي ظهور المساعدين الافتراضيين الذين يعملون بالصوت مثل خدمات أليكسا وإيكو سبوت.

يسعى جرار وغيره من علماء الكمبيوتر من جميع أنحاء الشرق الأوسط إلى تغيير هذا الوضع، إذ يعملون على توسيع نطاق شمولية عالم التكنولوجيا من خلال تحسين فهم الذكاء الاصطناعي للغة العربية بما يتجاوز اللغة العربيّة الفصحى. ويأمل هؤلاء العلماء أن تتمكّن البرامج والتطبيقات والخدمات الصوتية التي أصبحت شائعة بشكل متزايد من فهم لهجات اللغة العربية التي يقدّر عددها بثلاثين لهجة.

شبح الجهاز

يُعرف فرع الذكاء الاصطناعي الذي يسمح لأجهزة الكمبيوتر بمعالجة وتفسير اللغة البشرية باسم معالجة اللغات الطبيعية. وعندما نطلب من أليكسا، وهي المساعد الافتراضي من شركة أمازون الذي يعمل بالصوت، تشغيل أغنية، فهي تستخدم تقنيات معالجة اللغات الطبيعيّة لتحليل أوامرنا الصوتيّة. وتُستخدم هذه التقنية أيضًا من قبل أدوات الترجمة الآلية مثل ترجمة غوغل. وفي الواقع، ليس بالمفاجئ أن الطريقة التي تُجمّع بها أجهزة الكمبيوتر اللغات مختلفة عن الطريقة التي يستخدمها البشر.

شابان تونسيان يتحدثان في مقهى في سيدي بوزيد، لكن أجهزة الكمبيوتر والأجهزة الأخرى لا تستطيع دائمًا فهم اللهجات.

في هذا السياق، أفاد جرار أن "أجهزة الكمبيوتر تتعلم اللغات من خلال الإحصاءات" ومن أجل ترجمة لغة إلى أخرى "يجمع الكمبيوتر ملايين أو مليارات الجمل التي تحمل المعنى ذاته بكلا اللغتين المختلفتين، وبهذه الطريقة يُستنتج أي ترجمة هي الأكثر شيوعًا". علاوة على ذلك، ينسب الباحثون بعض الخصائص للكلمات، مثل موضعها في جملة أو سابقتها ولاحقتها، مما يقدّم مجموعة من البيانات التي يمكن للكمبيوتر أن يعتمد عليها في إحصائياته. وكلما زاد عدد البيانات المجمّعة، كانت أكثر دقة.

بمعنى آخر، تعدّ البيانات أهمّ عنصر عندما يتعلق الأمر بتلقين اللغات لجهاز الكمبيوتر. في المقابل، أشار جرار إلى أنه من الصعب تجميع ما يكفي من البيانات عندما يتعلق الأمر باللهجات العربية. وأورد جرار الذي يتخصّص في اللهجة الفلسطينيّة أنه "قبل عصر مواقع التواصل الاجتماعي لم تكن هناك أية لهجة مكتوبة فعلاً. وكان الأمر مقتصرا على الطريقة التي تتحدّث بها مع عائلتك وأصدقائك، وهو ما يعني أن العرب يكتبون كلماتهم على مواقع التواصل الاجتماعي مثلما يلفظونها صوتيًا".

قال جرار إن تحليل اللهجة الفلسطينية ينطوي على جمع أعداد هائلة من النصوص، ثم العمل على كل كلمة وإرفاقها بالخصائص أو القيم التي تعرّفها والتي ينبغي على الكمبيوتر أن يتعلمها

بدأت اللهجة العربية تكتب على مواقع الإنترنت بعد الإنجليزية والفرنسية والإسبانية التي تستخدم الأبجدية الرومانية. ويعني ذلك أن العلماء مثل جرار لديهم بيانات قليلة لتدريب الذكاء الاصطناعي مقارنة بزملائهم الذين يعملون على لغات أخرى. (على سبيل المقارنة، إذا أراد شخص ما العمل على مشروع باستخدام تقنيات معالجة اللغات الطبيعيّة باللغة الإنجليزية، فسيجد البيانات التي يحتاجها. وفي هذا الإطار، قال جرار: "يركّز الجميع على اللغة الإنجليزية، بالتالي لم تعد هذه اللغة مهمّة في هذا المجال").

في هذا الشأن، قال جرار إن تحليل اللهجة الفلسطينية ينطوي على جمع أعداد هائلة من النصوص، ثم العمل على كل كلمة وإرفاقها بالخصائص أو القيم التي تعرّفها والتي ينبغي على الكمبيوتر أن يتعلمها، مثل موضعها في الجملة أو سابقتها ولاحقتها أو معناها في اللغة الإنجليزية ومعناها في اللغة العربية الفصحى. ولكن كانت 2016 سنة إحراز التقدّم، إذ أفاد جرار أنه "أصبح بإمكان أجهزة الكمبيوتر في الوقت الراهن فهم اللهجة الفلسطينية". وعموما، كان جرار ثاني شخص يدرب الكمبيوتر بلغة عربية، بينما كان الجيش الأمريكي أوّل من حقق مثل هذا التقدّم بالنسبة للمصريين.

لغز اللغة العربية

لا يتأثر تدريس اللغة العربية على أجهزة الكمبيوتر بنقص البيانات فقط، فاللغة تضمّ كذلك العديد من السمات التي يمكن أن تزيد من درجة غموضها وصعوبتها. وفي هذا السياق، ذكر باحث معالجة اللغات الطبيعيّة المصري، علي فرغلي أن "اللغة العربية لا تشمل خاصيّة الحروف الكبيرة في أوّل الكلمات، وهي طريقة للإشارة إلى أسماء الأشخاص والأماكن والشركات. فضلا عن ذلك، تُغيّر الحروف العربية طريقة شكلها كلّما تغيّر موضعها في الكلمة". بالإضافة إلى ذلك، يمكن إنشاء كلمات أطول في اللغة العربية من خلال ربط عناصر أصغر من اللغة سوية.

مخطوطة القرآن بجامعة برمنغهام: الطبيعة المعقدة للغة العربية تعني أن أجهزة الكمبيوتر لا يمكن أن تتعلم العربيّة بالطريقة ذاتها التي تتعلم بها اللغة الإنجليزية واللغات الأخرى.

أشار فرغلي إلى أنه "بالإمكان أن تحتلّ إحدى الكلمات المعقدة وظيفة مبتدأ أو فعل أو مفعول به، وفي كثير من الأحيان يكون من الممكن تقسيم الكلمة المعقدة بثلاث طرق مختلفة أو أكثر، مما يزيد من الغموض". وعلى سبيل المثال، تتكوّن عبارة "them killed He" في اللغة الإنجليزية من ثلاث كلمات، ولكن تترجم هذه الجملة باللغة العربية في كلمة واحدة فقط: "قتلهُم".

علاوة على ذلك، يقدّم فرغلي مثالاً آخر، حيث أشار قائلا: "يمكن لكلمة باللغة العربية مثل "وَفيّ" أن

تُعتبر كلمة واحدة مشتقّة من صفة الوفاء، أو يمكن تقسيمها إلى كلمتين: "وَ"، و"في". وفي الحقيقة، إنّ يمكن تفكيك مثل هذه الكلمات بأكثر من طريقة تجعل من رفع الالتباس في اللغة العربية مهمّة شاقة في معالجة اللغات الطبيعية.

سبب ضُعف تمويل اللغة العربية

تُعدّ هذه المشاكل في التعلم الآلي شائعة، كما أنّ العلماء يحاولون معالجتها منذ أوائل الثمانينات. وفي الواقع، تسارع نسق الأبحاث بعد وقوع حدث حاسم. وفي هذا الصدد، قال فرغلي: "بعد الحادي عشر من أيلول/سبتمبر، موّلت الحكومة الأمريكية بسخاء الجامعات ومراكز البحوث والشركات الخاصة للعمل على معالجة اللغات الطبيعية في اللغة العربية".

في وقت سابق من هذا العام، أعلنت هيئة أبوظبي للإعلام، أبوظبي ميديا، أنها تعمل على تطوير أول روبوت لمذيعة إخبارية بتقنية الذكاء الاصطناعي ناطقة باللغة العربية في العالم

في السياق نفسه، أضاف فرغلي: "طبّق العلماء الأمريكيون أحدث التقنيات لتطوير أنظمة الترجمة الآلية العربية. وكان لهذه التقنيات تأثير إيجابي على عمل معالجة اللغات الطبيعية في اللغة العربية في العالم العربي". وعلى الرغم من هذه الطفرة، إلا أن اللغة العربية، واللهجات على وجه الخصوص، ظلت تعاني من نقص الموارد إلى حدّ ما، حيث تقوم شركات رئيسية مثل "أمازون" و"غوغل" و"آي بي إم" بتمويلها بصفة أقل من اللغات اللاتينية.

في هذا الشأن، يقول عبد الله فزع، وهو رائد أعمال في التكنولوجيا من الأردن، إن هذا النقص في الاستثمار يرجع إلى حد كبير إلى وجود حوافز أكبر لتطوير منتجات بلغات أخرى أكثر تداولا، مثل الماندرين، أو لديها المزيد من التطبيقات التجارية، مثل الإسبانية. وفي مقام أوّل، أنشأ فزع "أرابوت"، أحد أوائل روبوتات الدردشة باللغة العربية. ويتيح البرنامج للعملاء طرح أسئلة حول المنتجات عبر الإنترنت، التي يردّ عليها بعد ذلك من خلال الكمبيوتر.

في الموضوع نفسه، أضاف فزع أنه "إذا قمنا بإنشاء روبوت دردشة باللغة الإنجليزية، سيكون السوق أكبر بكثير بالنسبة له. وعموما، فيما يخصّ روبوتات الدردشة، فإن "آي بي إم" هي المنافس الرئيسي، ولكنها تعمل فقط على المستوى الأساسي بالنسبة للغة العربية".

يعمل فريق من الباحثين في الجامعة اللبنانية الأمريكية في لبنان على تحسين معالج اللغات الطبيعية باللغة العربية بحيث يمكن استخدامه لتحليل محتوى مواقع التواصل الاجتماعي لالتقاط الأحداث والمعلومات المهمة وغير المكتشفة.

من جهة أخرى، توجد مشاريع أخرى ذات جاذبية تجارية آخذة في الظهور. وفي وقت سابق من هذا العام، أعلنت هيئة أبوظبي للإعلام، أبوظبي ميديا، أنها تعمل على تطوير أول روبوت لمذيعة إخبارية بتقنية الذكاء الاصطناعي ناطقة باللغة العربية في العالم. فضلا عن ذلك، قالت شركة موضوع الأردنية، العام الماضي إنها بدأت العمل على مساعدة إفتراضية، مثل أليكسا أو سيري، تدعى سلمى ستعمل باللغة العربية وجميع لهجاتها.

الأمم المتحدة: أجهزة الكمبيوتر بحاجة إلى الفهم

لا تقتصر هذه التطورات على القطاع التجاري. وعلى سبيل المثال، يعمل فريق من الباحثين في الجامعة اللبنانية الأمريكية في لبنان على تحسين معالج اللغات الطبيعية باللغة العربية بحيث يمكن استخدامه لتحليل محتوى مواقع التواصل الاجتماعي لالتقاط الأحداث والمعلومات المهمة وغير المكتشفة.

في الوقت نفسه، يوضّح فادي زرقيت الذي يقود فريق الباحثين قائلا: "نحن نطور ونستخدم تقنيات

معالجة اللغات الطبيعية لتحليل النص العربي بما في ذلك مواقع التواصل الاجتماعي إنستغرام وسنابشات وفيسبوك وتويتر". وأردف زرقيت قائلا إنّه "يمكن استخدام أدوات التحليل التي نطورها للكشف عن إشارات الأشخاص والأماكن والعنف والشكاوى وغيرها من الأحداث على مواقع التواصل الاجتماعي. وعموما، يساعد ذلك على استكمال الصورة التي تعرضها وسائل الإعلام الرئيسية، عن طريق التقاط الأحداث التي قد لا يتفطّن إليها الإعلام".

يُعدّ فهم أجهزة الكمبيوتر للغة العربية أمرًا مهمًا للمشروع، لأنه يسمح بتحديد الأحداث التي لم يكتشفها المجتمع الدولي. من جهة أخرى، تسعى المنظمات غير الحكومية الدولية بدورها إلى إمكانية الاستثمار في هذا المجال. ومن جانبه، أفاد مارتن فيليش، المستشار السياسي في إدارة الشؤون السياسية وبناء السلام التابعة للأمم المتحدة، وهو جزء من فريق يعمل على تطوير نظام يستخدم أجهزة الكمبيوتر المدربة على اللهجات العربية لإجراء مجموعات تركيز جماعية. وعلى نحو خاص، يعمل النظام على طرح أسئلة على آلاف الأشخاص في مناطق النزاعات، ثم يتفحص الإجابات للعثور على نقاط مشتركة.

يشعر جرار بالتفاؤل إزاء مستقبل أجهزة الكمبيوتر واللغة العربية، ويفخر بالتقدم الذي أُحرز بالفعل

علاوة على ذلك، أوضح فيليش أنّ "معالجة اللغة الطبيعية بالعربية وبلغات أخرى تُعد بمثابة إشكال استمرّ لوقت طويل بالنسبة للشؤون العامة والسياسية، وذلك لأننا مهتمون بفهم اهتمامات الناس واحتياجاتهم بشكل أفضل، وهو ما يمكن أن يساعدنا على تصميم حوار وعمليات إحلال سلم أكثر استدامة". وفي الغالب، ستسمح هذه التقنية للأمم المتحدة بالحصول على إشارات تحدث في الوقت الحقيقي وتدلّ على المشاعر العامة والأحاسيس داخل المناطق المعزولة. كنتيجة لذلك، يُعتبر الحصول على نظام يعمل باللغة العربية أمرا بالغ الأهمية بالنظر إلى الصراعات في جميع أنحاء المنطقة.

يشعر جرار بالتفاؤل إزاء مستقبل أجهزة الكمبيوتر واللغة العربية، ويفخر بالتقدم الذي أُحرز بالفعل. وفي هذا الصدد، أورد جرار قائلا إن "معالجة اللغة الطبيعية للغة العربية أقوى بكثير مما كانت عليه قبل خمس أو 10 سنوات". ويشرح جرار أنه بمجرد الانتهاء من معالجة اللهجات، سيكون التحدي التالي هو العمل على أجهزة الكمبيوتر التي تفهم حقًا اللغة، بدلا من تخمين الترجمات استنادًا إلى الإحصائيات.

كما خلص جرار إلى القول إنه "إذا أخبرت جهاز كمبيوتر أنك في إجازة، فيمكنه ترجمة ذلك، ولكن إذا طرحت عليه السؤال التالي: إلى أين أذهب في إجازة؟، فلن يتمكن من تقديم الإجابة. كنتيجة لذلك، يمكن لجهاز الكمبيوتر معالجة المحتوى، لكنه لا يستطيع فهم معنى الكلمات، الأمر الذي ينبغي العمل عليه مستقبلا". بالإضافة إلى ذلك، تُعتبر اللغة الإنجليزية أكثر تطورا من اللغة العربية على الرغم من أنها لا تعدّ مثالية. في المقابل، أعتقد أنه في غضون بضع سنوات سيكون باستطاعة سيري تقديم إجابات باللغة العربية حول جميع أسئلتنا المطروحة".

المصدر: ميدل إيست آي

رابط المقال: https://www.noonpost.com/29339/