مجموعة تقنيات اللغة العربية:

التزام بالإبداع وجودة الأبحاث

معهد قطر لبحوث الحوسبة

يونيو 2023

البحث العلمي في تقنيات اللغة العربية في معهد قطر لبحوث الحوسبة

معهد قطر لبحوث الحوسبة (QCRI) هو معهد بحثي وطني، أنشئ عام 2010 من قبل مؤسسة قطر (QF) للتربية والعلوم وتنمية المجتمع كجزء من مهمة مؤسسة قطر لبناء قدرات الابتكار والتكنولوجيا في قطر.

يركز معهد قطر لبحوث الحوسبة على تحديات الحوسبة التي تتناول الأولويات الوطنية للنمو والتنمية، ويقوم بإجراء أبحاث علمية على مستوى عالمي متعددة التخصصات وذات صلة باحتياجات قطر والمنطقة العربية الأوسع والعالم. يتضمن ذلك أبحاثًا متطورة في تقنيات اللغة العربية تعمل عليها مجموعة تقنيات اللغة العربية (ALT).

لعبت مجموعة ALT على مدار العقد الماضي دورًا مهمًا في تطوير مجال معالجة اللغات الطبيعية (NLP) للغة العربية، ضمن مهمتها الأساسية في ضمان ازدهار اللغة العربية في العالم الرقمي. وكان هذا من خلال إجراء بحوث متقدمة في تقنيات اللغة العربية وإنتاج الأدوات وإنشاء الموارد اللغوية ودعم تطوير الحلول والخدمات للدول الناطقة باللغة العربية.

طورت مجموعة ALT قدرات لغوية تيسر طيفاً واسعًا من مقدرات معالجة وتحليل اللغة العربية، وواجهت تحديات اللغة العربية ولهجاتها لتبرز على مستوى الريادة العالمية في مجالات معالجة اللغة الطبيعية والتعرف على الكلام والترجمة الآلية والإجابة عن الأسئلة.

وتواصل مجموعة ALT العمل الدؤوب في البحث العلمي ذي الأهمية والتأثير وتستغل الفرص المتاحة لتسخير اكتشافاتها في الابتكار ومعالجة المشكلات المعاصرة.

أنشطة مجموعة ALT

تعاونت مجموعة ALT علمياً بشكل واسع مع المؤسسات الأكاديمية والشركات بما في ذلك معهد ماساتشوستس للتكنولوجيا وجامعة كارنيجي ميلون وشبكة الجزيرة وقطر ليفينج وشركة بوينج ومؤسسات القطاع العام القطري كالمجلس الأعلى للتعليم وسدرة الطب والمركز الاجتماعي والثقافي للمكفوفين. ومؤخراً في مجال التوعية تتعاون المجموعة مع منظمة الأمم المتحدة لتمكين نشر محتوى عالي الجودة غير متحيز عبر مواقع الويب الخاصة بالدول الأعضاء.

يخدم أعضاء مجموعة ALT دورياً في رئاسة المؤتمرات وورش العمل الرئيسية في مجالهم. وقد نظمت المجموعة في عام 2014 مؤتمر الأساليب التجريبية في معالجة اللغة الطبيعية (EMNLP) وهو أحد أهم المؤتمرات في هذا المجال، واستضافت عام 2023 ورشة العمل IEEE SLT2022 حول تقنيات اللغة المنطوقة وهو حدث ذو اعتبار وتقدير بين خبراء النطق الآلي.

وترحب مجموعة ALT بشكل مستمر بقدوم الباحثين والأكاديميين الزائرين الذين يرغبون في قضاء سنوات التفرغ في معهد قطر لبحوث الحوسبة.

تأثير مجموعة ALT

تحقق مجموعة ALT تأثيرها الأساسي من خلال:

- ابتكار ونشر تقنيات اللغة العربية
- العمل العلمي، بما في ذلك المنشورات الأكاديمية والأنظمة التوضيحية التطبيقية
- قيادة ودعم المجتمع البحثي
- نقل التكنولوجيا

معهد قطر لبحوث الحوسبة
مجمع البحوث لجامعة حمد بن خليفة
ص. ب. 5825، الدوحة، قطر
www.qcri.qa

- العمل التطوعي الذي يخدم قطر والمجتمع الدولي

ابتكار ونشر تقنيات اللغة العربية

توفر وثائقنا المقدمة وصفاً تفصيلياً للإنجازات التقنية الرئيسة:

1. فراسة Farasa - مقدرات معالجة اللغة الطبيعية للعربية الفصحى الحديثة (MSA) واللهجات العربية
2. أسد Asad - تحليلات دلالية لمحتوى وسائل التواصل الاجتماعي
3. تنبيه Tanbih- تجميع وتحليل الأخبار و اكتشاف الأخبار المزيفة
4. شاهين Shaheen - ترجمة آلية من العربية إلى الإنجليزية ((MT
5. QATS - تحويل الكلام العربي المنطوق إلى نص
6. ناطق NatiQ - تركيب الكلام المنطوق وتحويل النص إلى نطق (TTS) للعربية الفصحى واللهجات العربية
7. NeuroX - فهم النماذج اللغوية وتقييم جودتها.

تغطي هذه التقنيات القدرات الأساسية لدعم اللغة العربية (فراسة) والتي تيسر التحليلات الدلالية المتقدمة للمحتوى (أسد وتنبيه) وتدعم الترجمة الآلية للغة العربية (شاهين) لكل من الفصحى واللهجات الدارجة.

علاوة على ذلك، طورت مجموعة ALT قدرات متقدمة لإدراك الكلام العربي (QATS) والنطق الآلي عالي الجودة (.NatiQ)

تستخدم معظم هذه التقنيات شبكات التعلم العميق والتي تتسم بالفعالية العالية لكنها تفتقر إلى الشفافية. من أجل ذلك، تجري مجموعة ALT بحثًا علمياً لفهم هذه النماذج وتقييم جودتها (.NeuroX)

العمل العلمي

لمجموعة ALT حضور قوي في المناسبات الأكاديمية البارزة حيث تقدم نتائج البحوث والأنظمة التوضيحية وتنظم ورش العمل وتشارك بتصميم المهام المشتركة والتحديات البحثية للمجتمع العلمي لمواجهة تحديات خاصة.

ينشر الفريق بانتظام في المؤتمرات العريقة مثل ACL و EMNLP و NAACL و AAAI و ICLR و COLING و EACL و LREC و INTERSPEECH و ICASSP. ونظمت مجموعة ALT عام 2014 مؤتمر الأساليب التجريبية في معالجة اللغة الطبيعية (EMNLP) وهو أحد أهم المؤتمرات في مجاله، ومؤخراً استضافت المجموعة SLT2022، ورشة عمل IEEE حول تقنيات اللغة المنطوقة في يناير كانون الثاني 2023.

منشورات ALT في المؤتمرات

المؤتمر		المنشورات	العروض	ورش العمل
ACL	Association for Computational Linguistics	2)Shaheen) 4) NeuroX) 7) Tanbih) 1) Farasa) 1) QATS)	1) NeuroX)	1) NeuroX) 1) Shaheen)
NAACL	North American Chapter of the Association for Computational Linguistics	6) Shaheen) 3) NeuroX) 6) Tanbih) 3) Farasa)	1) Shaheen)	
EMNLP	Empirical Methods in Natural Language Processing	2) Shaheen) 4) NeuroX) 8) Tanbih)		1) NeuroX) 1) Shaheen)

		1) Farasa)		
COLING	International Conference on Computational Linguistics	Shaheen 2 NeuroX 1	1)Farasa)	
AAAI	Association for the Advancement of Artificial Intelligence	NeuroX 1	2) NeuroX)	
ICLR	International Conference on Learning Representations	NeuroX 2		
EACL	European Chapter of the Association for Computational Linguistics	Shaheen 2 NeuroX 1	Shaheen 1 NeuroX 1 Asad 1	
LREC	Language Resources and Evaluation Conference	Farasa 2 Asad 2		
IJCAI	International Joint Conferences on Artificial Intelligence	Tanbih 2		

بالإضافة إلى ذلك، تنشر مجموعة ALT كتباً وأبحاثاً في مجلات محكمة.

الإنجازات العلمية لكل مشروع

المشروع	المنشورات	العروض	ورش العمل
Farasa (Arabic NLP)	8 https://farasa.qcri.org/	1	6
Asad (Semantics)	3 https://asad.qcri.org/research	1	13
News)) Tanbih	+40 https://tanbih.qcri.org/publications/	5	10
Shaheen (Machine Translation)	29	3	3
QATS (Arabic Speech Recognition))			
(Text-to Speech) NatiQ		1	
Language) NeuroX Model Understanding)	25	4	2

قيادة ودعم المجتمع البحثي

وفرت مجموعة ALT وصولاً مجانيا إلى ما طورته من تقنيات اللغة على شكل حزم برمجيات وواجهات برمجية للأدوات، كما وفرت مجموعات البيانات لتمكين الاختبارات المعيارية وإمكانية تكرار التجارب بشكل علمي.

الوصول عبر الإنترنت والواجهات البرمجية وحزم الأدوات البرمجية

حزم الأدوات البرمجية (للتحميل)	واجهات برمجية	متاح عبر الإنترنت	المشروع
Yes	Yes	Yes	Farasa
No	Yes	Yes	ASAD
Yes	Yes	Yes	Tanbih
No	Yes	Yes	Shaheen
N/A	N/A	N/A	*QATS
No	Yes	Yes	NatiQ
Yes	No	Yes	NeuroX

* تم نقل تقنيات QATS إلى الشركة الناشئة Kanari بهدف تحويلها لمنتج تجاري

المجتمع العلمي المتخصص بالنطق

Arabicspeech.org – منصة لجمع بيانات النطق العربي

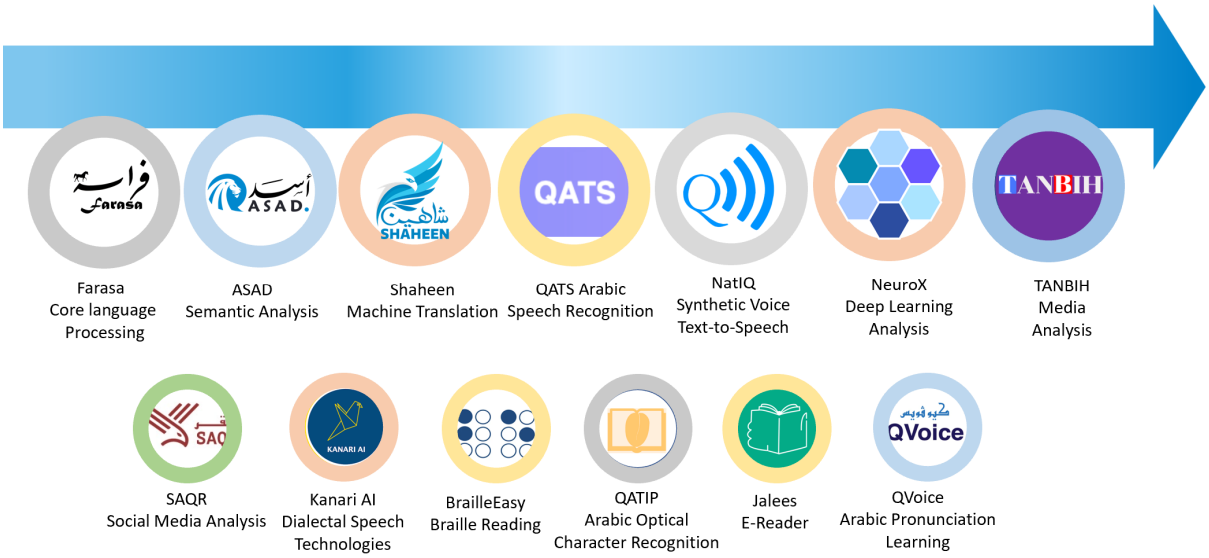
يعد توافر موارد البيانات عالية الجودة أمراً أساسياً لإنشاء تقنيات اللغة وتحسينها، وقد لعبت ALT دوراً قيادياً في مجتمع النطق العربي ونسقت الجهود التي تفيد علوم النطق العربي وتقنيات الكلام. تشمل البيانات المتاحة ما يلي:

- MGB-2: أكثر من 1200 ساعة تم جمعها من قناة الجزيرة، إلى جانب 130 مليون كلمة من Aljazeera.net. وقد تم إضافة النصوص يدوياً دون معلومات التوقيت.
- MGB-3: بهدف فهم الكلام باللهجة المصرية بشكل مفتوح. تم وسم كل جملة بالنص المنطوق من قبل أربعة أشخاص، وتم جمع أكثر من 15 ساعة من يوتيوب.
- MGB-5: بهدف التعرف على الكلام العربي المغربي بشكل مفتوح. 14 ساعة من YouTube مع نصوصها المكتوبة إلى جانب 90 ساعة مصنفة حسب النوع دون نصوص.
- QASR: هو أكبر مجموعة كلام عربية حتى اليوم بحوالي 2000 ساعة صوتية مع وسوم نصية تفصيلية ولهجات متعددة وعدة متحدثين في آن.
- تعدد المتحدثين - الإسكوا: تم جمعها على مدى يومين من اجتماعات لجنة الأمم المتحدة الاقتصادية والاجتماعية لغرب آسيا (الإسكوا) في عام 2019.
- مجموعة بيانات تعدد المتحدثين باللغة العربية الدارجة: تتضمن مجموعة البيانات باللهجة المصرية مع النصوص لمدة ساعتين من ADI-5 في تحدي MGB-3
- ADI-5: أكثر من 50 ساعة تم جمعها من قناة الجزيرة تتضمن أربع لهجات إقليمية: المصرية (EGY) والشامية (LAV) والخليج (GLF) وشمال إفريقيا (NOR) بالإضافة إلى العربية الفصحى الحديثة (MSA). مجموعة البيانات هذه هي جزء من تحدي MGB-3.

- Lexicon - oc: معجم النطق العربي القائم على الجرافيم

الابتكار ونقل التكنولوجيا

أنتج فريق ALT في السنين القليلة الماضية تقنيات مكنت البحث العلمي والتطبيقات من خلال إظهار تأثير تقنيات اللغة العربية في سياقات مختلفة. ويبين الشكل التالي التقدم المحرز في التقنية ومجالات التطبيق:

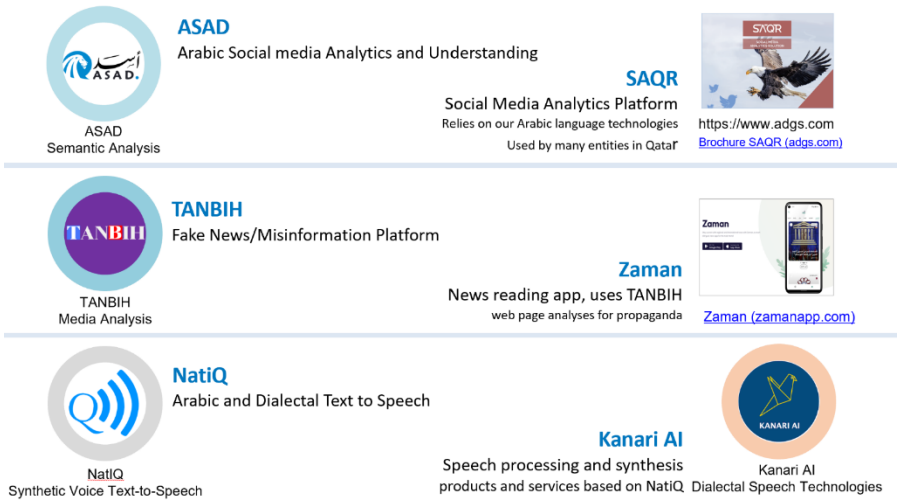


الشكل .1 تطور البحوث والابتكار في معهد قطر لبحوث الحوسبة.

نجح معهد قطر لبحوث الحوسبة في تحويل عدد من تقنيات المجموعة إلى منتجات من خلال جهود هندسية ضمن المعهد ومن خلال نقل التكنولوجيا إلى الشركات التي ابتكرت منتجات وخدمات تستعملها.

- من أسد Asad إلى صقر Saqr: طور الفريق الهندسي في معهد قطر لبحوث الحوسبة منصة مراقبة لوسائل التواصل الاجتماعي تباع الآن من خلال شريك موزع
- من تنبيه Tanbih إلى زمان Zaman: طور الفريق الهندسي في معهد قطر لبحوث الحوسبة تطبيقاً للهواتف لتجميع الأخبار يستهدف المنطقة العربية
- الشركة الناشئة كناري Kanari AI: قام معهد قطر لبحوث الحوسبة بنقل تقنيات QATS وNatiQ وأتاح شاهين Shaneen API إلى كناري لابتكار منتجات وخدمات تناسب السوق العربية.

الشكل .1 تطور البحوث والابتكار في معهد قطر لبحوث الحوسبة.



الشكل .2 نقل التكنولوجيا من البحث إلى التطبيق

أعمال المجموعة التطوعية لخدمة قطر والمجتمع الدولي.

لطالما كانت مجموعة ALT سباقة في تقديم قدراتها التقنية وخبراتها لمواجهة التحديات وتمكين الابتكار، وقد تم القيام بهذا العمل كخدمة للمجتمعات المحلية والدولية. على سبيل المثال:

- نظام كوفيد 19 ((ASAD
- تحليل المناهج التعليمية في المنطقة العربية ((Farasa
- تقديم المساعدة إلى منظمة الأمم المتحدة ((ASAD/Tanbih

الإشادة بإنجازات مجموعة ALT

تم الإشادة بأبحاث وتقنيات مجموعة ALT من حيث الجودة ومدى التأثير، وانعكس ذلك في العديد من الجوائز التي حصلت عليها المجموعة:

Prize	Description	Date
Best innovation Award	An automated transcription and translation system developed by QCRI won ARC'18 Best Innovation Award	March 2018
BBC Hackathon	Best in Show award for machine speech translation	2017

كان لأعضاء مجموعة ALT تأثير في زيادة الوعي حول تقنيات اللغة العربية ومساهمتها في تطوير اقتصاد المعرفة في قطر وجلب المنافع للمجتمعات الناطقة باللغة العربية حول العالم. وقد ظهرت المجموعة في بيانات صحفية ونشرت مقالات ذات صدى واسع.

تصريحات صحفيه:

الناشر	عنوان المقال	التاريخ

December 2017	Reading a Neural Network's Mind	MIT Press Release
December 2017	Reading a Neural Network's Mind	Science Blog
February 2019	Putting Neural Networks under the microscope	MIT Press
February 2019	Putting Neural Networks under the microscope	ACM Communications

مقالات علمية:

ARAB WORLD SPECIAL SECTION: BIG TRENDS.

Connecting Arabs: Bridging the Gap in Dialectal Speech Recognition, Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 124-129

By Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, Assal Alqudah

10.1145/3451150

Connecting Arabs: Bridging the Gap in Dialectal Speech Recognition | April 2021 | Communications of the ACM

ملاحظات ختامية

هدف هذه الوثيقة تقديم لمحة عامة عن البحث والابتكار في تقنيات اللغة العربية الذين يحققهما فريق ALT في معهد قطر لبحوث الحوسبة.

توفر وثائقنا المقدمة وصفاً تفصيلياً للإنجازات التقنية الرئيسة:

1. فراسة Farasa - مقدرات معالجة اللغة الطبيعية للعربية الفصحى الحديثة (MSA) واللهجات العربية
2. أسد Asad - تحليلات دلالية لمحتوى وسائل التواصل الاجتماعي
3. تنبيه Tanbih- تجميع وتحليل الأخبار و اكتشاف الأخبار المزيفة
4. شاهين Shaheen - ترجمة آلية من العربية إلى الإنجليزية ((MT
5. QATS - تحويل الكلام العربي المنطوق إلى نص
6. ناطق NatiQ - تركيب الكلام المنطوق وتحويل النص إلى نطق (TTS) للعربية الفصحى واللهجات العربية
7. NeuroX - فهم النماذج اللغوية وتقييم جودتها.

حيث نوضح التفاصيل الفنية لكل تقنية والتأثير العلمي والتأثير المجتمعي ونقل التكنولوجيا من خلال المنتجات والشركاء الذين يقدمون تقنيات اللغة العربية للمستخدمين.