نحو أداء أفضل لأنظمة التعرف على الكلام العربي المتصل بإستخدام القواعد التركيبة للغة العربية

د. ضياء أبوزينة
abuzeina@ppu.edu
كلية تكنولوجيا المعلومات وهندسة الحاسوب، جامعة بوليتكنك فلسطين، الخليل، فلسطين

تعتبر تقنية التعرف الآلي على الكلام واحدة من العناصر الرئيسة في تكوين أنظمة معالجة اللغات الطبيعية، بحيث يستطيع المستخدم إدخال بياناته الصوتية ليتم معالجتها وتقديم النتائج بشكل نصي. إلا أن هناك عدد من العوامل تحول دون الحصول على أنظمة تعرف على الكلام تنافس قدرة الإنسان، ومن هذه المعوقات ما هو خاص باللغة العربية مثل وجود التشكيل من عدمه بالإضافة إلى مسألة تعدد اللهجات، ومنها ما هو عام لجميع اللغات مثل ظاهرة التغيّر الصوتي. نقدم في هذ البحث نتائج دراسة تتعلق بتحسين الأداء في أنظمة التعرف الآلي على الكلام العربي المتصل. ومع أن نتائج هذه الدراسة لم تظهر ان الطريقة المقترحة تؤدي الى تحسن الأداء، إلا انه من المفيد عرضها وتقديم نتائجها للباحثين.

تعتمد الطريقة المقترحة على اعادة تقييم الفرضيات (N-best list rescoring) الناتجة في انظمة التعرف على الكلام، بحيث يتم ارجاع الفرضية الاكثر تطابقاً مع القواعد التركيبية للغة العربية. فبمجرد انتهاء عملية التعرف تخضع النتائج لعملية إعادة تقييم لاختيار الفرضية الأفضل من الناحية التركيبية. اما بخصوص القواعد التركيبية فيتم الحصول عليها من خلال خوارزميات التنقيب عن البيانات (Data mining) للمدونة (corpus) المستخدمة بعد وسمها (Tagging). فعلى سبيل المثال، تحتوي الجملة التالية على الكلمات وأوسامها حيث تم استخدام أداة الوسم (Stanford tagger) لوسم الكلمات علماً أن القراءة لهذه الجملة تتم من اليسار إلى اليمين حسب مخرجات (Stanford tagger).

➔ (DTNN/اليوم DTJJ/الأمريكية NNP/كيميكلز NNP/دال NN/وشركة DTNNP/السعودية NNP/أرامكو NN/شركة VBD/قالت)

يوضح الجدول 1 معنى عدد من الأوسام المستخدمة في نظام ستانفورد لوسم أنواع الكلمات. ففي الجملة السابقة فإن وسم كلمة "قالت" هو (VBD). حسب الجدول فإن معنى هذا الوسم هو (فعل ماضي).

الجدول **1**. أنواع الكلمات حسب نظام ستانفورد

#	الوسم	المعنى	أمثلة
…	…	…	…
8	JJ	صفة	جديدة، قيادية
9	JJR	صفة مقارنة	أدنى، كبرى
…	…	….	…
14	VBD	فعل ماضي	أصدرت
15	CD	رقم	مئة، ألفين
…	…	…	…

إن مخرجات الـ (Tagger) فتتمثل في مجموعة القواعد الأشهر التي تم استنباطها من مجموعة أوسام الجمل المدخلة. يوضح الجدول 2 عدد من القواعد التي تم استخلاصها. وتشير القاعدة الأولى إلى أن وسم الكلمة الخامسة يكون حرف جر إذا كان وسم الكلمة الرابعة رقم ووسم الكلمة السادسة اسم مبتدء بـ الـ وأن الدقة في هذه القاعدة بلغت اكثر من 95%.

الجدول **2**. عدد من القواعد التركيبية التي تم استخلاصها بناء على أوسام الكلمات

1	TAG4=CD TAG6=DTNN ==> TAG5=IN acc:(0.95635)
2	TAG1=VBD TAG3=DTJJ TAG7=DTNN ==> TAG2=DTNN acc:(0.95635)
3	TAG7=CD TAG8=IN ==> TAG9=DTNN acc:(0.95222)
…	…

استخدمت هذه القواعد لإعادة تقييم نتائج التعرف على الكلام من خلال إنتاج عدد من الفرضيات لكل جملة يتم فحصها (ميزة إنتاج عدد من الفرضيات متاحة في أنظمة التعرف على الكلام ومنها سفينكس). ويعاد تقييم الفرضيات من خلال إيجاد الفرضية التي تتناسب أكثر مع قواعد اللغة. فمن الممكن أن تكون الفرضية الثالثة متوافقة أكثر مع القواعد المستخلصة بشكل أكبر من الفرضية الأولى (الأعلى احتمالاً حسب ما هو مستخدم في أنظمة التعرف على الكلام) ففي هذه الحالة يتم اختيارها (أي الفرضية الثالثة) كأفضل نتيجة ممكنة ويتم بالتالي إرجاعها إلى المستخدم كنتيجة نهائية.

بالرغم من التحسن في عملية التعرف قد ظهرت في بعض الجمل، إلا انه لم يظهر لدينا أن هذه الطريقة تؤدي إلى تحسن ملحوظ في الأداء. ولعل عاملين رئيسين قد أثرا سلباً على نتائج هذه الطريقة. يتمثل العامل الأول في نظام ستانفورد للتعرف على أقسام الكلام اذ يحتوى على أخطاء في نتائجه. بينما يتمثل العامل الثاني في قضية التشكيل. إن استخراج الفرضيات للكلام المشكل يؤدي إلى إعطاء فرضيات يكون الاختلاف بينها على أساس حركات التشكيل وليس الكلمات التي تتعامل معها هذه الطريقة.

أمثلة على تحسن في الاداء في عملية التعرف على الكلام:

A waveform of a speech sentence with its text form	هذا وقد بلغت مبيعات شركة فورد موتورز في الصين خلال عام ألفين وخمسة	**A waveform of a speech sentence with its text form**	حذر البنك الدولي دول الخليج العربية من فخ المزيد من عائداتها النفطية في مشروعاتٍ
As recognized by the Baseline system	هذا وقد بلغت مبيعات شركة فورد موتورز التسعين خلال عام ألفين وخمسة	**As recognized by the Baseline system**	حذر البنك الدولي دول الخليج العربية من فخم المزيد من عائداتها النفطية في مشروعاتٍ
Found at ➔	**Hypothesis # 36**	**Found at ➔**	**Hypothesis # 50**
As recognized by the enhanced system	هذا وقد بلغت مبيعات شركة فورد موتورز في الصين خلال عام ألفين وخمسة	**As recognized by the enhanced system**	حذر البنك الدولي دول الخليج العربية من فخ المزيد من عائداتها النفطية في مشروعاتٍ