مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الموارد اللُّغويَّة الحاسوبيَّة

مباحث لغوية 01

تحرير
د. مُحسن رَشوان د. المُعتزّ بالله السَّعيد

الباحثون:
د. عبد العاطي هوَّاري
د. سامِح الأنصاري
د. المُعتزّ بالله السَّعيد
د. مُحسن رَشوان

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

مباحث لغوية ٥٦

الموارد اللُّغويَّة الحاسوبيَّة

تحرير

د. مُحســـن رَشـــوان
د. الـمُعتزّ بالله السَّعيـد

الباحثون:

د. عبد العـاطي هوَّاري
د. الـمُعتزّ بالله السَّعيـد
د. سامِح الأنصـــاري
د. مُحســـن رَشـــوان

١٤٤١هـ - ٢٠١٩م

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

الموارد اللُّغويَّة الحاسوبيَّة

الطبعة الأولى
١٤٤١ هـ - ٢٠١٩ م
جميع الحقوق محفوظة
المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٤١هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
رشوان، محسن
الموارد اللغوية الحاسوبية./ محسن رشوان؛ المعتز بالله السعيد .-الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٩ -٥٤ -٨٢٢١ -٦٠٣ -٩٧٨
١ - اللغة العربية - معالجة البيانات أ. السعيد ، المعتز بالله (مؤلف مشارك) ب. العنوان
ديوي ٤١٠,٢٨٥ ١٠١٦٩/ ١٤٤٠
رقم الإيداع: ١٠١٦٩/ ١٤٤٠
ردمك: ٩ -٥٤ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج

دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com

لايسمح بإعادة إصدار هذا الكتاب، أو نقله في أي شكل أو وسيلة،
سواء أكان إلكترونية أم يدوية أم ميكانيكية، بما في ذلك جميع أنواع تصوير المستندات بالنسخ، أو التسجيل أو التخزين، أو أنظمة الاسترجاع، دون إذن خطي من المركز بذلك.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

فهرس الكتاب

الموضوع	الصفحة
كلمة المركز	٧
مقدمة	٩
الفصل الأول : الموارِد المعجَمِيَّة العَرَبِيَّة الحاسُوبيَّة	**١١**
١- مدخل إلى الموارِد المعجَمِيَّة العَرَبِيَّة الحاسُوبيَّة	١٣
٢- في التعريف بالموارِد المعجَمِيَّة الحاسوبية	١٣
٣- الموارِد المعجَمِيَّة ومعالجة اللغات الطبيعية	٢٢
٤- الصناعة المعجمية الحاسوبية	٢٦
٥- الموارِد المعجَمِيَّة العربية الحاسوبية	٣٧
٦- الأفكار البحثية المقترحة في إطار العمل المعجمي الحاسوبي العربي	٤٢
الفصل الثَّاني: المدَوَّنات اللُّغَوِيَّة	**٥١**
١- في مفهوم المُدَوَّنات اللُّغَوِيَّة	٥٣
٢- إرهاصات المنهج، وتطَوُّر دراسة المدَوَّنات اللُّغَوِيَّة	٥٥
٣- المُدَوَّنات اللُّغَوِيَّة العربيَّة	٥٨
٤- أنواع المُدَوَّنات اللُّغَوِيَّة	٦١
٥- عنونة/ تذييل المدَوَّنات اللُّغَوِيَّة	٦٦

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

٦- المُدَوَّنات اللُّغَوِيَّة وآليَّة فهرسة النُّصُوص	٧٤
٧- مجالات الإفادة من المدَوَّنات اللُّغَوِيَّة	٧٧
٨- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة	٨٣
٩- مِن المواقع الإلكترونيَّة التَّعليميَّة والإرشاديَّة	٨٨
الفصل الثَّالث: الشَّبكات الدِّلاليَّة	**٩٣**
١- التحليل الدلالي للجملة: لمحة تاريخية	٩٥
٢- لغة الشبكات الدلالية الحاسوبية العالمية	٩٧
٣- المكونات اللغوية للغة الشبكات الدلالية الحاسوبية العالمية	٩٩
٤- موارد وأدوات لغة الشبكات الدلالية الحاسوبية العالمية	١٠٦
٥- تطبيقات المعالجة الآلية للدلالة باستخدام لغة الشبكات الدلالية الحاسوبية العالمية	١٢٤
٦- دعوة للمشاركة	١٢٨
الفصل الرابع: موارد التَّعلُّم الآلي (مدخل إلى التَّعلُّم الآليّ)	**١٣٥**
١- شجرة القرار	١٣٧
٢- مصنِّف بايز المبسط	١٣٨
٣- الشبكات العصبية	١٤٠
٤- آليَّات المتجهات الداعمة (Support Vector Machines -SVM)	١٤٥
٥- نماذج ماركوف المُخَبَّأة (Hidden Markov Models - HMMs)	١٤٨
الفصل الخامس: نمذَجة اللُّغة	**١٦٥**
١- النحو العدديّ (N-gram)	١٦٨
٢- التنعيم (Smoothing)	١٧٥
٣- موضوعات تساعد على تحسين النَّحو العدديّ	١٨٢
٤- تقويم قوة النَّحو العَدَدِيّ	١٨٣
٥- أمثلة على مجالات الإفادة من النَّحو العَدَدِيّ	١٨٥
٦- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة	١٨٥
الباحثون	١٨٩

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الفصل الخامس
نمذَجة اللُّغة

د. مُحسِن رَشوان

١- النَّحو العدديّ.
٢- التَّنعيم.
٣- موضوعات تساعد على تحسين النَّحو العدديّ.
٤- تقويم قوة النَّحو العَدَدِيّ.
٥- مجالات الإفادة من النَّحو العَدَدِيّ.
٦- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

تمهيد

اللُّغات الحية في الحقيقة معقَّدة بما فيه الكفاية لتلبية حاجة الإنسان في التَّعبير عن مشاعره وأفكاره المتجدِّدة، إذ يمكن للإنسان أن يعبِّرَ عن معنى يجول في خاطره بعددٍ كبير جدًا من الجُمَل الَّتي تؤدِّي نفسَ المعنى. وربما تختلف عن بعضها في الدِّقة والبلاغة، والمشاعر المحيطة بالمعنى... إلخ. وهذا يجعل وضعَ إطار رياضيٍّ دقيقٍ للتعبير عن فهم المتحدث وقصده أمرًا بالِغَ الصُّعُوبة – إن لم يكُن مُستَحيلاً – في الوقت الحالي؛ وفي الوقت ذاته لا نستطيعُ الاستغناءَ عن نمذجة اللُّغة الَّتي تُوَجِّهُ التِّقنيات اللُّغَوِيَّة إلى تحقيق أهدافها المنشودة في مجالاتٍ مُتَعَدِّدة، كالتَّعرُّف الآليّ على الكلام المكتوب أو المنطوق.

وعلى سبيل المثال، في مجال التَّعَرُّف على الكلام المنطوق، لو افتَرَضـــنا أنَّ المتَحـدِّثَ نطقَ جملةً تحتوي على ١٠٠ فُونيما متتاليًا (حوالي ١٠–١٢ كلمة متصلة) – آخذينَ في الاعتبارِ أن أدقَّ الأنظمة التي تتعرَّف على الكلام المنطوق لا يتجـاوز متوسط دقَّتها ٨٠٪ لكل فونيم على حدة – فإنَّ درجـــة دِقَّة التّقنية على مستـــوى الجمل إذا تخلَّينا عن استخدام النَّموذَج اللُّغَوِيّ ستكونُ على النَّحو المبَيَّن في الجدول التَّالي:

عدد الفونيمات المكونة للكلمة أو الجملة بافتراض متوسط دقة ٨٠٪ لكل فونيم	
عدد الفونيمات	دقَّة التَّعَرُّف
١	٨٠٪
٢	٦٤٪ = ٢^(٨٠٪)
٣	٥١,٢ =٣^(٨٠٪)
١٠ (متوسط الكلمة)	١٠,٧٪ = ١٠^(٨٠٪)
٥٠ (جملة قصيرة)	٠,٠٠١٤٪ = ٥٠^(٨٠٪)
١٠٠ (جملة متوسطة)	٠٪ ≈ ١٠٠^(٨٠٪)

الجدول ٥–١: دقَّة الكلمات والجمل بدون نموذج لغويّ.

ووفقًا لهذه النَّتائج، سيُؤَدِّي الاستغناء عن النَّموذَج اللُّغَوِيّ إلى نتائج ليست ذات قيمة، وبالتَّالي ستُصبِحُ تقنية التَّعَرُّف على الكلام المنطوق عديمة الفائدة بخُلُوِّها من هذا النَّموذَج. أمَّا إذا اعتمدنا عليه فإنَّ الكلمات العربية في صُورتها المفردة ستتحرَّك

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

من ٧, ١٠٪ إلى أكثر من ٩٠٪ (في ظروف تسجيل مناسبة)، بمساعدة النَّماذِج اللُّغَوِيَّة باعتبارِها مجموعة من المعلومات الرِّياضِيَّة الموضوعة في قالبٍ رياضِيّ؛ وبعبارةٍ أخرى، تُساعد «نمذجة اللغة» (Language Modeling) في تحقيق الفائدة من تقنيات اللُّغات.

ونستطيعُ التَّمثيل على ذلك بتحليل المقطع الصَّوتيّ «ذَهَبَ إلى»، حيث تحتملُ اللَّفظة «إلى» أن تكونَ ١. «**آلي**»، أو ٢. «**آلا**»، أو ٣. «**إلى**». ونستطيعُ أن نستَدِلَّ على الاحتمال الأقرب إلى الصَّواب بتحليل تتابُع هذه الكلمات في سياقاتها اللُّغَوِيَّة؛ وبافتراض أنَّنا قُمنا بتحليل كلمة «**ذَهَبَ**» وتعرَّفنا عليها بشكلٍ صحيح، فإنَّنا سنجد أنَّ كلمة «**إلى**» هي الأكثرُ التصاقًا بها، مايعني أنَّ الاحتمالَ الثَّالثَ أقربُ إلى الصَّواب.

١ - النحو العدديّ (N-gram)

هناك بعضُ الطُّرُق الَّتي تُستَخدَم في توجيه تقنيات اللُّغة وتطبيقاتها إلى الاحتمال الأقرب إلى الصَّواب من النَّاحية اللُّغَوِيَّة؛ ويُعَدُّ النَّحو العدديّ «N-gram» أوسعَ هذه الطُّرُق انتشارًا وأكثرها استخدامًا. وسنُحاول الوُقُوفَ -فيما يلي- على أهمِّيَّة النَّحو العدديّ ودوره في تقنيات اللُّغة، مُقَدِّمِينَ له بالحديث عن الاحتمالات [٤،٥، ١٤].

١, ١ - حساب الاحتمالات والاحتمالات الشَّرطيَّة

إذا كانت لدينا مُدوَّنةٌ لُغَوِيَّةٌ تضُمُّ مليون كلمة، وكانت إحدى كلماتها قد ورَدَت ١٠٠٠ مرة، فإنَّنا نستطيعُ تحديدَ احتمال ورُود هذه الكلمة في وثيقة تتشابهُ مادَّتُها مع مادَّة المدونة اللُّغَوِيَّة باستخدام المعادَلة التَّالية:

$$\text{احتمال الورود} = P(w) = \frac{\text{عدد مرات ورود الكلمة في المدونة}}{\text{عدد كلمات المدونة كلها}}$$

$$P(w) = \frac{١٠٠٠}{١٠٠٠,٠٠٠} = ٠,٠٠١ = ٠,١٪$$

أي واحد في الألف.

حيث ترمُز w إلى الكلمة، (اختصارًا لـ Word)،

وترمز $P(w)$ إلى احتمال الوُرُود، (اختصارًا لـ Probability of Word).

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

دعنا نعرف الاحتمالات المشروطة في هذا المثال: ورد في القرآن الكريم كله عدد ٧٧٩٣٤ كلمة، وكانت كلمة «الله» هي الأكثر ورودًا فيه، وجاءت هذه اللفظة الكريمة ٢٧٠٧ مرة، فكانت مرفوعة في ٩٨٠ مرة، ومنصوبة في ٥٩٢ مرة، ومجرورة في ١١٣٥ مرة. فلو سألنا عن احتمال ورود كلمة الله في القرآن الكريم كله ستكون الإجابة:

$$P(\text{الله}) = \frac{٢٧٠٧}{٧٧٩٣٤} = ٪٣,٤٧$$

بينما لو سألنا عن كلمة القرآن مرفوعة في القرآن كله تكون الإجابة:

$$P(\text{اللهُ}) = \frac{٩٨٠}{٧٧٩٣٤} = ٪١,٢٦$$

ماذا لو سألنا هذا السؤال: ما احتمال ورود كلمة (الله) مرفوعة منسوبة إلى كل كلمات (الله) في القرآن الكريم؟ أو بعبارة أخرى: ما احتمال ورود كلمة (الله) مرفوعة بشرط نسبها إلى كلمة (الله) في القرآن كله؟ سيكون التَّعبير رياضيًّا على هذا النَّحو:

$$P(\text{اللهُ} / \text{الله})$$

وتعني الشَّرطة المائلة في التَّعبير الرِّياضيّ السَّابق أنَّ شرطَ حساب احتمال ورود كلمة (الله) مرفوعة هو ورود كلمة (الله) أيًّا كان تشكيلها. ويكون حسابها كالآتي:

$$P(\text{اللهُ} / \text{الله}) = \frac{٩٨٠}{٢٧٠٧} = ٪٣٦,٢$$

١,٢ - النَّحو العدديّ الأحاديّ (Uni-gram)

بعد أن قدَّمنا فكرةَ الاحتمالات الشَّرطيَّة، يُمكنُنا أن نُقَدِّمَ فكرةً عن النَّحو العدديّ. كما أسلفنا في المقدمة أنَّنا في حاجة ماسَّة إلى معلومات عن اللُّغة وعن تردُّد كلماتها وترابطها معًا، لنتمكَّنَ من دعم الحل الصحيح في تقنيات كثيرة من تقنيات اللُّغات الحية. والواقعُ أنَّ ظُهُور النَّحو العدديّ الاحتماليّ في الخمسينيَّات من القرن العشرين باعتباره مسارًا إحصائِيًّا يُستَخدَم في مُعالجَة اللُّغات الحيَّة قد لاقى عزوفًا من قبل اللُّغَوِيِّين في ذلك الوقت نتيجةَ ما أفرزَته نظريَّات اللُّغَوِيّ الأمريكيّ نَعُوم تشومسكي من نقدٍ لهذا المسار. ولكن بعد أن نجحت شركة IBM في السبعينيَّات من العودة إلى

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

النَّحو العدديّ بنجاح، اتَّجَهَ الباحثون في تقنيات اللُّغات الحية إلى الاستعانة به، حتى غدا أساسا لا غنى عنه لمطوري هذه التقنيات.

دعنا نأخذ مثالاً مطولاً لفهم النَّحو العدديّ [أو الإحصائِيّ] N-gram. لو تصوَّرنا أنَّ لدينا مدونةً لغوية مُبَسَّطة تتكون من هاتَين الجملتَين:

«ذهب محمد إلى المدرسة»
«حين وصل محمد إلى المدرسة قابل زميله أحمد»

عدد الكلمات في هذه المدونة المبسطة ١٢ كلمة؛ وبإضافة رمز لبداية جملة (مَرَّتين)، ورمز نهاية جملة (مَرَّتين)، وكأنهما كلمتان مضافتان لمفردات المدونة، يكون عدد الكلمات ١٦ كلمة. أي: عدد مفردات المدونة ١٦ مفردة (١٢ + بدايتين لجملتين ونهايتين).

وقبل الشُّروع في توضيح مفهوم النَّحو العددِيّ، نودُّ أن نُشِيرَ إلى قيام عالِمِ الرِّياضيَّات الرُّوسيّ أندريه ماركوف (١٨٥٦-١٩٢٢) بوضع نموذجٍ رياضيٍّ مبسطٍ للتَّنبُّؤ بالمستقبل بالاستعانة فقط ببضع خطوات من الماضي. وسوف نستفيد من تبسيطه الرياضيّ فيما يلي:

لنحسب للمدونة المبَسَّطة السَّابقة (والَّتي لا يتعدَّى محتواها ١٦ كلمة) حسابات تدخل في مفهوم النَّحو العدديّ:

أولاً: تُعرَفُ الدَّرجة الأولى في النحو العدديّ بـ «النحو الأحادي uni-gram» أو 1-gram؛ وفيه نحسب فقط احتمالية تكرار كل كلمة بصرف النظر عن ما قبلها أو ما بعدها، على النَّحو المبَيَّن في الجدول التَّالي:

م	مُفردات المدونة	التَّرَدُّد (الوُرُود)	النَّحو الأحاديّ	النِّسبة
١	بداية جملة	٢	٢/١٦ = ١/٨	٠,١٢٥
٢	ذهب	١	١/١٦	٠,٠٦٢٥
٣	محمد	٢	٢/١٦ = ١/٨	٠,١٢٥
٤	إلى	٢	٢/١٦ = ١/٨	٠,١٢٥
٥	المدرسة	٢	٢/١٦ = ١/٨	٠,١٢٥

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

م	مُفردات المدونة	التَّرَدُّد (الوُرُود)	النَّحو الأحاديّ	النِّسبة
٦	حين	١	١٦/١	٠,٠٦٢٥
٧	وصل	١	١٦/١	٠,٠٦٢٥
٨	قابل	١	١٦/١	٠,٠٦٢٥
٩	زميله	١	١٦/١	٠,٠٦٢٥
١٠	أحمد	١	١٦/١	٠,٠٦٢٥
١١	نهاية الجملة	٢	١٦/٢ = ٨/١	٠,١٢٥
	المجموع	١٦	١,٠٠	١,٠٠

الجدول ٥-٢: حسابات النَّحو الأحاديّ لمفردات المَدَوَّنة المبَسَّطة.

١ ,٣ - النَّحو العدديّ الثنائيّ (Bi-gram)

يمكن الارتقاء درجةً وحساب النَّحو العدديّ إذا نظرنا خلفنا لكلمة واحدة، واستعنَّا بهذه المعلومة لحسابات المستقبل. فبالنَّظر إلى الجدول رقم (٥-٢) سنُلاحظُ أنَّنا نضَع في حساباتنا (بداية الجملة) و (نهاية الجملة). ويسمَّى هذا بالنَّحو الثُّنائيّ [٥، ٢٤].

		الكلمة السابقة										
		بداية جملة	ذهب	محمد	إلى	المدرسة	حين	وصل	قابل	زميله	أحمد	نهاية جملة
١	بداية جملة											
٢	ذهب	١										
٣	محمد		١					١				
٤	إلى			٢								
٥	المدرسة				٢							
٦	حين	١										
٧	وصل						١					
٨	قابل					١						

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

		الكلمة السابقة										
		بداية جملة	ذهب	محمد	إلى	المدرسة	حين	وصل	قابل	زميله	أحمد	نهاية جملة
٩	زميله								١			
١٠	أحمد									١		
١١	نهاية جملة					١					١	
	المجموع	٢	١	٢	٢	٢	١	١	١	١	١	

الجدول ٥-٣: النحو الثنائي للمدَوَّنة.

الاحتمال الشرطى للنحو الثنائي $P(w_n/w_{n-1})$	عدد ورود الكلمة $C(w_{n-1})$	عدد ورود الكلمتين معاً $C(w_n,w_{n-1})$	$\frac{C(w_n,w_{n-1})}{C(w_{n-1})}$	$P^*(w_n/w_{n-1}) = \frac{C(w_n,w_{n-1})+0.01}{C(w_{n-1})+121*0.01}$
P(ذهب/ بداية جُملة)	٢	١	٠,٥	٠,٣١٥
P(محمد/ ذهب)	١	١	١	٠,٤٥٧
P(إلى/ محمد)	٢	٢	١	٠,٦٢٦
P(المدرسة/ إلى)	٢	٢	١	٠,٦٢٦
P(نهاية جُملة/ المدرسة)	٢	١	٠,٥	٠,٣١٥
P(حين/ بداية)	٢	١	٠,٥	٠,٣١٥
P(وصل/ حين)	١	١	١	٠,٤٥٧
P(قابل/ المدرسة)	٢	١	٠,٥	٠,٣١٥
P(زميله/ قابل)	١	١	١	٠,٤٥٧
P(أحمد/ زميله)	١	١	١	٠,٤٥٧
P(نهاية جُملة/ أحمد)	١	١	١	٠,٤٥٧
اى تتابع ثنائى لم يرد عاليه	لو كانت: $C(w_{n-1})=1$	٠	٠	٠,٠٠٤٥
	لو كانت: $C(w_{n-1})=2$	٠	٠	٠,٠٠٣١

الجدول ٥-٤: النحو الثنائي للمدَوَّنة. العمود الثالث محسوب فيه النحو الثنائي بدون مراعاة

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

لـ OOV، والعمود الأخير محسوب فيه النحو الثلاثي بعد مراعاة OOV.

وهكذا، لو أردنا درجةً أخرى من نحوٍ أعمق فبإمكاننا أن نلجـأ للنَّحو الثُّلاثيّ (3-gram)، وعندئذ يكون مثلاً:

$$P(\text{محمد، إلى / المدرسة}) = \frac{٢}{٢} = ١$$

وعليه، يمكنُ حسابُ النَّحو الرُّباعيّ والخُماسيّ... إلخ.

والآن، نريد أن نقف عند مشكلة خطيرة في هذا الطرح، ألا وهي: ماذا نفعل مع الكلمات الَّتي لم ترد في سياق المدَوَّنة؟ سيكون احتمال وُرُودِها صفرًا، وهذا يتغير كثيرا إذا حسبنا أن ما لم نره في المدونة يكون احتمال وُرُودِه صفرا [٣، ٢١].

مثال: إذا قابلتنا عبارة (ذهب أحمد إلى المدرسة)، وأردنا الاستفادة من المدونة السابقة في استنباط نتائج مفيدة:

$$P(\text{أحمد .ذهب}) = ٠$$

سوف نجد أنها تساوي صفرا لأننا في الواقع لم نر هذا التركيب في المدونة التي استنبطنا منها نحونا الثُّنائيّ. ولو لم نجد حلاً لهذه المشكلة فإنَّ هذا سوف يسبب ضررًا بالغا لأيّ استخدامٍ لهذه النتائج، إذ إنَّ جملةً محتملة بصورة كبيرة، وربما بدرجة احتمال جملة (ذهب محمد إلى المدرسة)، لن نجدَ لها ما يدعمها من النَّحو الثُّنائيّ؛ والسَّبب أن المدوَّنات اللُّغَوِيَّة مهما كبرت فلن تغني عن أن واقع اللغات الحية متدفِّق ومتنامي. ويكادُ تعداد جملُ وتعبيرات هذا الواقع أن يكون لا نهائيًّا؛ فكيف نستنبط ما لم نره في المدونة؟

١ , ٤ – مشكلة: من خارج مُفردات المدَوَّنة (Out Of Vocabulary - OOV)

لو لم نجب على هذا السؤال ما أمكَنَ للنَّحو الإحصائيّ أن يكون مفيدا، لأنَّ ضرره سيكون أكبر من نفعه في كثير من الأحيان. وبعبارةٍ أخرى، لو لم يتمكَّن الباحثون من إيجاد حلول لهذه المشكلة لما كانت لهذا النحو قائمة.

تعالوا نفتَرض أن لدينا نحوًا فيه ١١ كلمة فقط، ووجدنا فيه ١١ حالة للنَّحو الثُّنائيّ يمكن أن نُقَدِّرَها تقريبا بـ ١١×١١=١٢١، أي أن هناك ١٢١ أيّ كلمة من مفردات

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

المدونة عقب كلمة أخرى (ويمكن أن تتكرر الكلمة، في مثل قوله تعالى: «وَجَاءَ رَبُّكَ وَالمَلَكُ صَفًّا صَفًّا»)، ولكن ورود ١٢ حالة فقط (كما فى الجدول رقم ٨- ٤) معناها أن هناك احتمالاً لـ ١٠٩ حالات لم ترد في المدونة. والحقيقة قد يكون ورود بعض التتابعات مستحيلا مثل ورود بداية جملة تتبعها بداية أو نهاية جملة .. إلخ، ولكن في مدونة حقيقية كبيرة لا يكون لهذه الاحتمالات أثر اذا اهملناها. وكذلك في الواقع الحقيقى يمكن أن نفرض أنه لن نرى إلا ~ ٥٠٪ من تتابع الكلمات بالنسبة لكل التتابعات الممكنة، هذا مقبول وحينئذ يمكن أن نحسب حساباتنا على توقع OOV ← ~ ٦٠ كلمة فقط. ولكن في مدونتنا البسيطة سنفرض للسهولة أن كل التتابعات ممكنة.

حل المشكلة:

لجأ كثير من الباحثين إلى محاولة تقدير احتمالات للمفردات والتَّتابُعات (الثنائية والثلاثية... إلخ) التي لم ترد في المدونة مع إعادة حساب التتابعات التي وردت بحيث يكون مجموع الاحتمالات واحدا صحيحا، لأن هذه من مسلمات نظرية الاحتمالات.

تعالوا نفترض أننا أضفنا مقدارًا ثابتًا، وقدره «٠,٠١» ، إلى كل احتمالات تتابع الكلمات؛ سوف نحتاج إلى إضافة ١٢١ مرة «٠,٠١» إلى البسط في ١٢١ حالة، شاهدنا فقط ١٢ حالة والباقي سنكتفي باعتبار وروده «٠,٠١» مرة تقديرا. ولذلك ستتغير الاحتمالات كما هو مبين في جدول رقم (٨-٤) العمود الأخير.

لنختبر نتائجَنا حتى الآن؛ هب أنَّنا سمعنا جملة، واختلط الأمر علينا بين جملتين:

- **«ذهب أحمد إلى المدرسة»**.
- **«قابل إلى أحمد زميله»**

(لنرى معًا كيف يُستخدم النَّحو العدديّ لترجيح أقرب الحُلُول إلى الصَّواب).

بتطبيق نظرية الاحتمالات:

$$P(\text{ذهب أحمد إلى المدرسة}) \approx P_*(\text{ذهب}) * P_*(\text{ذهب/أحمد}) * P_*(\text{أحمد/إلى}) * P_*(\text{إلى/المدرسة})$$

$$= \underset{(OOV)}{٠.٠٠٤٥} * \underset{(OOV)}{٠.٠٠٤٥} * ٠.٤٥٧ = ٥.٧٨ * ١٠^{-٧} = ٠.٠٠٦٢٥$$

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

بينما «**قابل إلى أحمد زميله**»

$$P(\text{قابل إلى أحمد زميله}) \approx P(\text{قابل}) * P(\text{قابل/إلى}) * P(\text{إلى/أحمد}) * P(\text{أحمد/زميله}) * P$$
$$= \frac{٠.٠٠٤٥}{(OOV)} * \frac{٠.٠٠٣١}{(OOV)} * \frac{٠.٠٠٤٥}{(OOV)} = ٣.٩٢ * ١٠^{-٩}$$

$$< P(\text{ذهب أحمد إلى المدرسة})$$

إذن: تكون الجملة الأولى هي المرَجَّحة.

٢- التنعيم (Smoothing)

تعالوا نعالج هذه المشكلة (من خارج المفردات) بطريقة أكثر عمقًا، تُعرَف بعملية التنعيم؛ أي: تنعيم قيم الاحتمالات الناتجة عن الحساب المباشر الناتج عن قسمة عدد التكرارات (سواء للكلمة أو الكلمتين المتجاورتين... إلخ) على العدد الكلِّيّ للكلمات في المدونة. وهناك طرق كثيرة للتنعيم نتعرف على أهمها.

٢, ١- التنعيم بالخصم (Smoothing by Discount)

كما أسلفنا فإن مشكلة عدم ورود كل الاحتمالات الممكنة في اللغة في قواعد البيانات المستخدمة في التدريب يسبب فشلا ذريعًا لاستخدام النحو العددي اذا لم تعالج هذه المشكلة. وهناك العديد من الطرق لتقدير هذه الاحتمالات.

- **تنعيم لابلاس (Laplace Smoothing)**

وتعتمد هذه الطريقة على تقدير عدد المرات التي نراها، ثم إضافة واحد لكل الحالات التي مرت بنا (بما في ذلك المرات التي مرت «صفر» مرة)؛ وبلغة الإحصاء:

$$P(w_j) = \frac{C_j}{N}$$

حيث C_j عدد مرات ورود الكلمة w_j ، و N العدد الكلي لكلمات المدونة.

وتصبح بعد طريقة تنعيم لابلاس:

$$P_{Laplace}(w_j) = \frac{C_j + 1}{N + V}$$

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

حيث V عدد المفردات المختلفة التي يمكن أن نصادفها. ولنأخذ مثالا لذلك:

هب أننا نملك مدونة بها ٣ كلمات فقط، وقدَّرنا أن هناك كلمة واحدة يمكن إضافتها؛ إذن ستكون (V=٤). وبافتراض ورود الكلمات كالآتي:

	عدد ورود الكلمة قبل التنعيم	عدد ورود الكلمة بعد التنعيم
$C_1=C(w_1)$	٣	٤
$C_2=C(w_2)$	٢	٣
$C_3=C(w_3)$	١	٢
$C_4=C(w_4)$	٠	١
	$V_1= 3, N_1=6$	$V_2= 4, N_2=10$

قبل التنعيم: N_1 في هذه الحالة = ٦، و V_2 = ٣ (مفردات):

$$P(w_1) = \frac{C_1}{N_1} = \frac{3}{6} = 0.5$$

$$P(w_2) = \frac{C_2}{N_1} = \frac{2}{6} = 0.33$$

$$P(w_3) = \frac{C_3}{N_1} = \frac{1}{6} = 0.167$$

لتصبح بعد تنعيم لابلاس $N_1 \leftarrow N_2$ = ١٠، و $V_1 \leftarrow V_2$ = ٤ (مفردات):

$$P_{Laplace}(w_1) = \frac{3+1}{6+4} = 0.4$$

$$P_{Laplace}(w_2) = \frac{2+1}{6+4} = 0.3$$

$$P_{Laplace}(w_3) = \frac{1+1}{6+4} = 0.2$$

$$P_{Laplace}(w_4) = \frac{0+1}{6+4} = 0.1$$

وإذا جمعت كل الاحتمالات الآن سوف تجد أنها تساوي الواحد الصحيح، بما يتفق مع إحدى مسلمات نظرية الاحتمالات.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

من الواضح أن إضافة واحد صحيح لكل مرات ورود المفردات يضعف بشكل ملموس احتمال المفردات التي وردت في المدونة بتكرار قليل بالنسبة لتلك المفردات التي لم ترد على الإطلاق؛ لذلك فإن هناك محاولات لتحسين هذا النوع من التنعيم بإضافة كمية ثابتة أقل من الواحد، وهذا يعتمد على حجم المدونة المستخدمة للتدريب.

ولكن كيف يتم تقدير عدد المفردات V؟ بالنسبة للنحو الأحادي، يتم تقديره على أساس المعرفة باللغة؛ ولكن اللغة العربية غنية جدا في عدد كلماتها؛ ففي مدونة من حوالى ١٥٨ مليون كلمة من الأخبار وجدنا بها ٩٥٠ ألف مفردة مختلفة بعضها عن بعض (كتاب، والكتاب، تُحسبان كلمتين مختلفتين)، وفي مدونة قريبة من ٦٦٠ مليون كلمة وجدنا قريبًا من ١,٨ مليون مفردة، لكنَّها احتوت على كمية كبيرة من الأخطاء اللغوية. فنحن نُقَدِّر المفردات الصحيحة في هذه الحالة بنحو ١,٤ مليون كلمة. لذلك عند التعامل مع مجال مثل الأخبار (وبالمناسبة، هو من المجالات الغنية بالمفردات لكثرة مجالاته الفرعية من سياسة واقتصاد، ورياضة، وعلوم، وحالات الطقس... إلخ) يمكن فرض أن عدد المفردات التي نتعامل معها قد يصل إلى أكثر من ٢ مليون مفردة، مع مُلاحظة أن اسم قرية جديدة أو مدينة وقَعَ بها زلزالٌ يضيف مفردة جديدة للمجال كل يوم.

ملاحظة: ليس بالضرورة أن تكون إضافة ١ هو الحل الوحيد المتاحإذ يمكن إضافة كمية ثابتة أقل – كما في المثال الذي سقناه آنفًا (وإن لم يكُن بالضرورة منخفضًا جدًّا كما فعلنا، إنما اخترنا القيمة القليلة (٠,٠١) لتناسب بساطة المدونة المستخدمة). وعادة ما يتم ذلك عبر عدة تجارب.

ومن الجدير بالذِّكر أننا في مثل هذه التجارب نحتاج إلى تقسيم المدونة إلى ٣ أقسام:

- القسم الأول للتعلم (في حالتنا لتعلم النحو العدديّ).
- القسم الثاني لاختيار أفضل القيم لبعض المعاملات (في حالتنا لاختيار أفضل قيمة للثابت (١، ٠,٥، ٠,١، ٠,٠١...).
- القسم الثالث للاختبار النهائيّ، ولا يجوز تغيير المعاملات ثم إعادة التجربة، لأن ذلك يعني أننا استعملنا قسم الاختبار في التدريب. لتوضيح ذلك، هب أننا أعددنا اختبارا

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

للطلاب فوجدنا مستواهم ضعيفًا في موضوع ما، فراجعناه معهم ثم أعدنا لهم نفس الامتحان! هذا لا يفرز الطالب الحافظ من الطالب الفاهم، لهذا الغرض خُصِّصَ القسم الثاني لاغراض ضبط متغيرات الحل.

- **خصم جود-تيورينج (Good-Turing Discount)**

وهي نظريَّةٌ إحصائيَّة، تُنسَب إلى العالمين «إرفنج جود (Irving John Good) وألان تيورينج (Alan Turing)» . وتعتمد منهجية الخصم هنا على فكرة بسيطة. إذا حسبنا عدد المفردات التي وردَت في المدونة مرة واحدة، ولنسمها N_1 ، وعدد المفردات التي وردَت في المدونة مرتَين، ولنسمها N_2، وهكذا سنحصل على N_3 ، N_4 ، N_5،....

وكذلك يمكن تقدير N_2 أي.. المفردات التي لم ترد في المدونة – ولو تقديرًا نظريًّا؛ فإننا لو افترضنا في تخصص معين أننا لن نتجاوز المليون مفردة، فإن

$$N_0=1,000,000 - N_1 - N_2 - N_3 - N_4 \ldots\ldots$$

ونعود لمنهجية تقدير احتمالات ورود المفردات:

$$C^* = (C + 1)\frac{N_{C+1}}{N_C}$$

حيث C هو عدد التكرارات الحقيقي، و *C هو التكرار التقديري لأغراض تنعيم الاحتمالات. ومن الملاحَظ في أيَّة مدونة أنه كلما زادت تكرارات المفردات كُلَّما قَلَّت أعدادها؛ وهذا يعني أن $N_c > N_{c+1}$ (هذه العلامة > تعني أن شمالها أكبر من يمينها). ولذلك يمكننا اعتبار أن $\frac{N_{C+1}}{N_C}$ هي مقدار التخفيض في الأعداد. ولو لاحظت أننا زدنا «١» وخفضنا بمقدار $\frac{N_{C+1}}{N_C}$ فستكون النتيجة:

- تخفيض في قيم الاحتمالات لما ورد من مفردات المدونة.
- وجود قيمة لاحتمالات ورود المفردات التي لم ترد في المدونة.

تعالَ نستدعي مدونتنا الصغيرة مرة أخرى:

- **ذهب محمد إلى المدرسة.**

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

- **حين وصل محمد إلى المدرسة قابل زميله أحمد.**

في مدونتنا السابقة؛ كما ورد منها في مدونتنا البسيطة ١١ كلمة.

إذن: تكون الأعداد N_c كالآتي:

N_0 (عدد المفردات التي لم ترد في المدونة) = ١٢١ − ١١ = ١٠٩

N_1 (عدد المفردات التي وردت مرة واحدة) = ٦

N_2 (عدد المفردات التي وردت مرتين) = ٥

محمد، إلى، المدرسة، بداية الجملة ونهاية الجملة، ولا تنس أن عدد الكلمات الكلِّيّ المشاهد في المدونة هو ١٦ كلمة = N . وعليه، سيكون تطبيق منهجية التنعيم باستخدام جود-تيورينج في تقدير احتمال النحو الثنائى الذى لم نره في المدونة:

$$P_{GT} \text{ (لأي تتابع لم نره)} = \frac{(0 + 1)\frac{N_1}{N_0}}{N} = \frac{6/109}{16} = 0.00344$$

والرمز $P_{GT}(x)$ يعني احتمال ورود (x) بتنعيم جود-تيورينج.

٢ , ٢- التنعيم باستخدام الإدراج (Interpolation)

ترتكز طرق التنعيم بالخصم على تقدير قدر مناسب من الاحتمالات للحالات التي لم نر فيها خصمًا مما ورد علينا في المدونة. ولكن التنعيم بالإدراج يفيد في حسن تقدير ما ورد علينا في المدونة، وذلك كالآتي؛ إذا أردنا تحسينا للنحو الثلاثي مثلا:

$$\hat{P}(w_n / w_{n-1}\ w_{n-2}) = \lambda_1\ P(w_n / w_{n-1}\ w_{n-2}) + \lambda_2\ P(w_n / w_{n-1}) + \lambda_3\ P(w_n)$$

بحيث يكون

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

وقيم λ_j أعلاهُ يمكن إيجادها بإجراء التجارب ووضع قيم مختلفة لها واختيار القيم التي تعطي أفضل النتائج للنحو (حسب المشكلة المستخدم فيها النحو).

ولقراءة هذه المعادلة لتكون مفهومة أكثر سنعيد كتابتها بالكلام:

الاحتمال المقدر للكلمة ن بشرط ورود الكلمتان ن-١، ن-٢=

ثابت يقدر من مدونة التدريب * احتمال كلمة ن بشرط ورود الكلمتان ن-١، ن-٢قبلها

ثابت آخر يقدر من مدونة التدريب * احتمال كلمة ن بشرط ورود كلمة ن-١ قبلها

ثابت آخر يقدر من مدونة التدريب * احتمال كلمة ن (أى النحو العددى)

على أن مجموع الثوابت الثلاثة لابد أن يكون واحدا صحيحا.

والفكرة من وراء هذا التحسين لتقدير النحو العددي تتبين من هذا المثال:

«**قال الله تعالى**» و «**رضي الله عنه**».

نفترض جدلاً أننا عند دراسة النحو الثلاثي لكلمتي (تعالى)، (وعنه) وجدنا أن تكرارهما متساوٍ في المدونة؛ ولكن كان ورود (الله تعالى) أكثر من (الله عنه)؛ وعليه.. فسيُساهم هذا في رفع احتمال (قال الله تعالى) عن (رضي الله عنه).

٢ ,٣- التنعيم بالتراجع (Smoothing using back-off)

- **تراجع كاتز (Katz back-off)**

يُستَخدَم تراجُع كاتز - عادةً - كمُكمِّلٍ لخصم جود-تيورينج؛ وتُستَوحى فكرتُه من التنعيم بالإدراج؛ ويمكن من خلاله فهم كيفية تقدير النحو العددي من درجة أعلى بدلالة النحو العددي من الدرجة الأدنى منه مباشرةً في المدونة.

لو أن عندنا نحوا ثلاثيا مطلوب تقديره، لأننا لم نره في المدونة، فإنَّ الطرق السابقة للخصم - وربما أفضلها حتى الآن جود-تيورينج -ستعطي كُلَّ ما لم نره نفسَ الاحتمال، ولكن «كاتز» يُقدِّرها اعتمادا على النحو الثنائي والأحادي إذا لزم الأمر. وهذا يعني أن

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

نعطي احتمالا أكبر للنحو الثلاثي المقدر لكلمات لم ترد في المدونة إذا كان نحوها الثنائي أكبر. ولبسط التعريف الرياضي انظر اسفل الصفحة[١].

- **التنعيم باستخدام طريقة «نزر-ناي» (Kenser-Ney)**

نستطيعُ الوُقُوفَ على هذه الطَّريقة من خلال المثال التَّالي:

أردت أن **أقرأ فأخرجت**..... ولم يرد في المدونة مثل هذه الجملة قط

بافتراض وجود كلمتين مرجحتين ولهما نفس النحو العددي الأقل، هما:

- «**النظارة**» (ما ورد في المدونة: عملت النظارة، وقعت النظارة، استخدمت النظارة، وضعت النظارة،....).
- «**بور**» والتي لم ترد إلا في (بور سعيد، بور فؤاد).

فإن كلمة «النظارة» تُرجَّح، لأن ورودها مع كلمات أكثر في المدونة يجعلها مرشحة للورود أكثر من كلمة «بور» فيما لم نره[٢].

١- نحتاج أن نعّرف:

أي تكرارات "x" $C(x)=$ count of x

احتمال x بعد الخصم (باستخدام طريقة من طرق الخصم السابقة) $P^*(x)$

وبدلا من استخدام w_{n-2} , $w_{n-1}w_n$ فإننا سنستخدم x,y,z لتكون المعادلات كالآتي:

$$P_{katz}(z/x,y) = \begin{cases} P^*(z/x,y) & if\ C(x,y,z)>0 \\ \propto(x,y)P_{katz}(z/y), & else\ C(x,y)>0 \\ P^*(z), & otherwise \end{cases}$$

$$P_{katz}(z/y) = \begin{cases} P^*(z/y), & if\ C(y,z)>0 \\ \propto(y)P^*(z), & otherwise \end{cases}$$

حيث $\propto$ تعني معامل التطبيع (لتجعل مجموع الاحتمالات ١ صحيحًا)، ولنقل اعتماد النَّحو العدديّ من درجة أعلى إلى درجة أقل. أما (x,y) أو (y) $\propto$ فتعني أن هذا المعامل متغير يعتمد على ما بين الأقواس.

وتجدُرُ الإشارةُ إلى أنَّ تراجع كاتز يمكن تعميمه على أيّ درجة من النحو العددي؛ أي أنَّ اقتصار المعادلات التي ذكرناها على النحو العددي من الدرجة الثالثة هو لمجرد التبسيط وتوضيح الفكرة. كذلك فإنَّ المعاملات $\propto$ يجري حسابها أيضًا من تكرارات النحو الأحادي والنَّحو الثنائي ... إلخ.

١- وتصاغ معادلاته كالآتي:

$$P(w_i/w_{i-1}) = \frac{C(w_{i-1}w_i)-d}{C(w_{i-1})} + \beta(w_i)\frac{|\{w_{i-1}:C(w_{i-1}w_i)>0\}|}{\sum_{w_i}|\{w_{i-1}:C(w_{i-1}w_i)>0\}|}$$

حيث d ثابت يطرح من كل احتمال لنحو ثنائيّ ورد في المدونة.

و $\beta(w_i)$ تُختار (وهي مختلفة من كلمة لأخرى) لتجعل مجموع الاحتمالات ١ صحيحًا.

و $|\{w_{i-1}:C(w_{i-1}w_i)>0\}|$ تعني عدد الكلمات المختلفة (w_{i-1}) التي ترد فيها مع في المدونة، مع ملاحظة أننا نحصي التنوع وليس عدد مرات الورود. مثال: لو وردت الكلمة (w_i) ١٠ مرات مع كلمة و٥ مرات مع كلمة أخرى فقط، فيكون مفهوم التعبير الرياضيّ المذكور هو ٢ وليس ١٥. (حيث يشير التعبير الرياضيّ |....| إلى أنَّ القيمة المذكورة تشمل عدد الأنواع، وليس عدد التَّكرارات).

و $\sum_{w_i}|\{w_{i-1}:C(w_{i-1}w_i)>0\}|$ تعني مجموع عدد المرات التي وردت فيها كلمات مختلفة في المدونة كلها.

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

٣- موضوعات تساعد على تحسين النَّحو العدديّ

هناك بعض الجوانب الَّتي تساعد على تحسين التقدير، ومن ذلك:

٣, ١ - النحو العددي الفئوي (Class Based N-gram)

خذ هذه الأمثلة:

كان راتب سعيد ١٠٠٠ جنيه في الشهر

ذهب على إلى الإسكندرية يوم الأربعاء

ركبت مريم طائرة مصر للطيران

فلو ارتبط النحو العددي برقم (١٠٠٠) فقط لما استفدنا من هذه المعلومة لو جاء الرَّاتب مختلفا في موضع جديد؛ ولكن يمكن أن نحدد أن هناك فئة من الأرقام يمكن أن يحل أحدها مكان الآخر. وكذلك أيام الأسبوع أو الشهور أو أسماء شركات الطيران... إلخ.

ففي المدونات قليلة العدد يمكن تعظيم الفائدة منها إذا عالجنا بعض الأسماء والأرقام باستخدام اسم الفئة التي تنتمي إليها هذه الأسماء أو الأرقام.

٣, ٢ - النحو العدديّ الموضوعيّ (Topic Based N-gram)

تتأثر النتائج كثيرا بشكل إيجابي إذا استخدمنا نحوًا عدديًّا من مدونة ذات موضوعات مشابهة للموضوع الذي نحن بصدده.

لذلك يمكن حساب النحو العددي لمدونات تحتوي كلٌّ منها على موضوعات متشابهة، مثل (مدونة سياسية، اقتصادية، علمية، قانونية،... إلخ). وهناك إضافات نوعية قد تكون مفيدة عند استخدام النحو العددي، ومنها الاستفادة من ظاهرة: الاستدعاء.

٣, ٣- دعم النَّحو العدديّ بالاستفادة من ظاهرة الاستدعاء

خذ هذا المثال:

ذهب إلى
ذهب محمد إلى
ذهب محمد وعلي إلى
ذهب محمد وعلي وسمير إلى

تلاحظ أن كلمة «ذهب» استدعت وجود كلمة «إلى» في كثير من الأحيان بعدها.

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

٤,٣ – النَّحو العدديّ متغير الطول (Variable length N-gram)

للنَّحو العدديّ أهمية قصوى في تطبيقات كثيرة؛ ولذلك نحتاج إلى دعمه بنظريات جديدة لغوية المنشأ مستوعبة لاحتياج الحاسوبيين، وخاصة مع اللغة العربية الَّتي تتمتَّعُ بظاهِرَتَي الاشتقاق والتَّوليد. وإذا كنَّا نحتاج في كثير من التطبيقات، مثل: التعرف على الكلام المنطوق في اللغة الإنجليزية، إلى ٦٤ ألف كلمة تغطي ٩٩٪ من احتياجات الكلمات في مجال معين (مثل مجال الأعمال Business) فإننا نحتاج إلى أكثر من ٦٠٠ ألف كلمة عربية لنقرب من درجة التغطية ٩٩٪. إن ذلك يجعل احتياجنا لمدونات كبيرة جدًّا لا مفر منه، والاحتياج إلى المعالجات اللغوية المسبقة ضرورة. ومن هذه المعالجات التحليل الصرفي لمعرفة السوابق واللواحق وجذع الكلمة، وربما نحتاج أيضا للوزن والجذر. (اللافت للانتباه أن العربية مبنية بعدد محدود من السوابق واللواحق والأوزان والجذور) إلا أن بناء النحو من هذه اللبنات له تحدياته ويستغرق جهودًا علميَّةً عميقةً من اللغويين والحاسوبيين للخروج بنحو عددي يستفيد من ميزات اللغة العربية وتطورها الصرفيّ، ويلبي حاجة التطبيقات المختلفة.

٤– تقويم قوة النَّحو العَدَدِيّ

نحتاج إلى تقويم كفاءة النحو المستخدم، ففي بعض التطبيقات يقيسون هذه الكفاءة بما يسمى مقدار «الالتباس» (Perplexity). وكلما قل الالتباس يعني ذلك كفاءة أعلى للنحو المستخدم. ويحسب مقدار الالتباس كما في المثال التالي:

على سبيل المثال، في اللغة الإنجليزية يحسب الالتباس عندما لا يكون هناك نحو على الإطلاق في تقنية التَّعَرُّف على الكلام المنطوق لعدد كلماتٍ مُحتَمَلة تدرَّبَت عليها التقنية، ومقدارها ٢٠٠٠٠ كلمة.

فكان مقدار الالتباس كما هو مُبَيَّنٌ في الجدول الآتي:

النحو العدديّ (N-gram)	الالتباس (Perplexity)
بدون نحو على الاطلاق	٢٠٠٠٠
النحو الأحادي (Uni-gram)	٩٦٢

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

النحو العدديّ (N-gram)	الالتباس (Perplexity)
النحو الثنائي (Bi-gram)	١٧٠
النحو الثلاثي (Tri-gram)	١٠٩

لننظر كيف انخفض مقدار الالتباس من ٢٠٠٠٠ بدون أي معلومات معطاة للنظام عن اللغة، إلى فقط ١٠٩ بعد استخدام النَّحو الثُّلاثِيّ. يمكن النَّظر إلى هذه الأرقام كالآتي: كأنّ المهمة التي تلقى على عاتق النظام قبل إعطاءه أي معلومات لغوية عند التعرف على الكلمة التي سمعها هي مهمة اختيار كلمة من ٢٠٠٠٠ كلمة. وليس له دليل على هذه الكلمة إلا ما يسمعه من صوت. وتنخفض درجة الالتباس لنفس المهمة إذا أفدنا النظام بمعلومات عن اللغة واستخداماتها وتتابعات كلماتها ملخصة في النحو الثلاثي لتصبح المهمة كما لو كانت هي التعرف على كلمة من ١٠٩ كلمة فقط باستخدام المعلومات الواردة من الصوت. هل نستطيعُ تصوُّرَ النتائج في الحالتين؟ الحالة الأولى: يفشل النظام تماما في الوصول إلى نتيجة لها أي اعتبار، أما في الحالة الثانية فإن النتائج يمكن أن تزيد عن ٩٠٪ كنسبة دقَّة في التعرف على الكلام المنطوق في ظروف مناسبة.

فبالرغم من بساطة فكرة النحو العدديّ إلا أنه – وبعد المعالجات المختلفة لما لم يره من كلمات وتتابعات – أصبحَ مُفيدًا للغاية وعمليا إلى درجةٍ كبيرة.

هل لديك أخي الباحث فكرة نيرة كهذه يصلح مع تطبيقها أن نصل لنتائج أفضل؟ إذا أمكن تمثيلُ اللُّغة رياضيًّا، فإنَّنا كعاملين في مجال تقنيات اللُّغة سنستفيد كثيرا من ذلك. **فشمِّر واجتهِد.**

وهناك العديد من الأعمال الآن في مجال توليد نماذج لغوية من الشبكات العصبية؛ والنتائج تتحدث عن تفوق ملموس عن النحو العددي، إلا أنها تحتاج لحسابات تأخذ في الغالب وقتا أطول بكثير من ذلك الوقت الذي يحتاجه النحو العددي.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

٥- أمثلة على مجالات الإفادة من النَّحو العَدَدِيّ

١- **التَّعرُّف على الكلام المنطوق**؛ كما أسلفنا. فربما كان هذا هو التطبيق الأول الذي أظهر قوة النَّحو العدديّ وتمَّ من خلاله علاج أخطر مُشكلاته، وهي عدم رؤيته لحالات كثيرة محتمَلة.

٢- **التَّدقيق الإملائيّ**؛ ولعلَّنا نُلاحظُ إشارات الخطأ الحمراء الَّتي يُنَبِّهُنا إليها البرنامج المكتبيّ «ميكروسوفت ورد MS-Word» ، وما يُرفقه من احتمالاتٍ للصَّواب. إن أصل العمليَّات التي يقوم بها هذا **المدقق الإملائي** هي من مثل النَّحو العدديّ.

٣- **الترجمة الآلية**؛ فقد تطورت **نُظُم الترجمة الآلية**، وأمكنَ من خلالها توليد عباراتٍ أكثر دقَّة عند استخدام النَّحو العدديّ في توليد الترجمة للُّغــة المستهــدفة.

٤- كما أن هناك في ساحة **محركات البحث** فرصة لتحسين البحث باستخدام النحو العدديّ.

٥- وكذلك في التطبيقات التعليمية **لتعليم اللغات** حيث يُستخدم الحاسب لتحليل ما كتبه المتعلم والحكم عليه. وهنا أيضا يستفاد من النحو العدديّ.

٦- هناك **نظم للتعرف على الحروف العربية**، فمنها المصمم للتعرف على الكلام المطبوع، ومنها المصمم للتعرف على الكلام المكتوب باليد، ولولا استخدامُ النحو العدديّ في هذه التَّطبيقات لكانت النتائج جد هزيلة ...

٦- أفكارٌ بحثيَّة لأطرُوحاتٍ علميَّةٍ مُستَقبليَّة

١- تكوين مدونة لبعض المجالات، تُختار موضوعاتها بحيث تحقق أعلى تغطية للكلمات التي يمكن أن تأتي في هذا المجال.

٢- البحث في أفكار جديدة لمعالجة مشكلة الكلمات التي لم نرها من قبل (في المدونة المخصصة للتدريب)، والتي نسميها التنعيم. كلما استفدنا من خصائص اللغة كلما كانت الحلول أوفق وأفضل.

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

٣- تكوين موارِد لغوية تساعد على تفضيل كلمة عن كلمة أخرى متقاربتين في النطق أو الكتابة اعتمادًا على خصائص دلالية للكلمتين.

٤- عمل معاجم مستنبطة من مدونات ترجح استخدام كلمة عن كلمة متقاربة معها فى الرسم أو النطق تبعا للسياق.

٥- تحتاج كثير من التطبيقات كالتعرف على الكلام المنطوق إلى معرفة نطق الكلمة الصحيح من سياقها - فوضع منظومة من القواعد المساعدة لضبط الكلمة من سياقها سيساعد كثيرا.

ببليوجرافيا مرجعيَّة

1. Bellegarda, J. R. & Monz, C. (2016). State of the art in statistical methods for language and speech processing. Computer Speech & Language, Elsevier, Vol. 35, pp. 163-184.
2. Cui, J. (2011). Integrating Linguistic and Statistical Knowledge in Language Modeling. BiblioBazaar.
3. Deng, L.; Liu, Y. (2018). Deep Learning in Natural Language Processing. Springer.
4. Farghaly, A. A. S. (2010). Arabic Computational Linguistics. University of Chicago Press.
5. Franz, A. & Brants, T. (2006). "All Our N-gram are Belong to You". Google Research Blog.
6. Friedenthal, S. & Moore, A. & Steiner, R. (2011). A Practical Guide to SysML: The Systems Modeling Language. Elsevier.
7. Goutte, C. (2009): Learning Machine Translation. MIT Press.
8. Huang, G. & Huang, G.B. & Song, S. & You, K. (2015). Trends in extreme learning machines: a review. Neural Networks, Elsevier, Vol. 61, pp. 32-48.

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

9. Johnson, M. & Khudanpur, S. P. & Ostendorf, M. & Rosenfeldm R. (2004). Mathematical Foundations of Speech and Language Processing. Springer.

10. Jozefowicz, R. & Vinyals, O. & Schuster, M. & Shazeer, N. & Wu & Y. (2016). Exploring the limits of language modeling. Cornell University.

11. Jurafsky, D. & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

12. Koren, B. & Vuik, K. (2010). Advanced Computational Methods in Sci-ence and Engineering. Springer.

13. Kumarm E. (2011). Natural Language Processing. I. K. International Pvt Ltd.

14. Luong, M. T.& Le, Q. V. & Sutskever, I. & Vinyals, O. & Kaiser, L. (2015). Multi-task sequence to sequence learning. Cornell University.

15. Lv, Y. & Zhai, C. (2009). Positional Language Models for Information Retrieval, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR).

16. Manning, C. D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing, MIT Press: ISBN 0-262-13360-1.

17. Matsumoto, Y. & Sproat, R. & Wong, K. & Zhang, M. (2006). Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead: 21st International Conference, ICCPOL 2006, Singapore, December 17-19, 2006, Proceedings.

18. Mishra, S. (2018). Artificial Intelligence and Natural Language Processing. Cambridge Scholars Publisher.

هذه الطبعة إهداء من المركز
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

19. Mulder, W. D. & Bethard, S. & Moens, M. F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. Computer Speech & Language, Elsevier, Vol. 30, pp. 61-98.

20. Neamat El, G.; Yee, S. (2018). Computational Linguistics, Speech and Image Processing for Arabic Language. World Scientific.

21. Olive, J. (2011). Handbook of Natural Language Processing and Machine Translation. Springer.

22. Sauro, J. & Lewis, J. R. (2012). Quantifying the User Experience: Practical Statistics for User Research. Elsevier.

23. Schimek, M. G. (2012). Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons.

24. Soudi, A. (2012). Challenges for Arabic Machine Translation. John Benjamins Publishing.

25. Srinivasa-Desikan, V. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing.

26. Su, Y. (2011). Knowledge Integration into Language Models: A Random Forest Approach. BiblioBazaar.

27. Sundermeyer, M. & Ney, H. & R Schlüter, R. (2015). From feed-forward to recurrent LSTM neural networks for language modeling. IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, pp. 517-529.

28. Wei, X. (2007). Topic Models in Information Retrieval. ProQuest.

29. Weinert, H. L. (2013). Fast Compact Algorithms and Software for Spline Smoothing. Howard L. Weinert.

30. Zhai, C. (2009): Statistical Language Models for Information Retrieval. Morgan & Claypool Publishers.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الباحثون

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الدكتور/ مُحسِن عبد الرَّازق علي رشوان

يشغل منصب أستاذٍ بقسم الإلكترونيات والاتصالات الكهربائية في كُلِّيَّة الهندسة – جامعة القاهرة. تخرَّجَ عام ١٩٧٧ وكان الأول على دفعته، وحصلَ على ثلاثة ماجستيرات، ثم على الدُّكتوراه من جامعة كوين بكندا؛ أشرف على أكثر من مائة رسالة ماجستير ودكتوراه. يدير الشَّركة الهندسيَّة لتطوير النُّظُم الرَّقمِيَّة RDI المتخَصِّصة في مجال تقنيات اللُّغة العربيَّة.

الدكتور/ المُعتزّ بالله السَّعيد طه

أستاذ الدِّراسات اللُّغويَّة الـمُساعد بجامعة القاهرة، وأستاذ اللِّسانيَّات الحاسوبيَّة المُشارك بمعهد الدَّوحة للدِّراسات العُليا، ومُنسِّق وَحدة الموارد الـمُعجميَّة بمشروع مُعجم الدَّوحة. نَشَرَ نحوَ ثلاثينَ ورقة علميَّةٍ، بالإضافة إلى عددٍ من الكتب في الـمُعجَمِيَّة العَرَبيَّة والدِّراسات اللُّغَوِيَّة الـمُعاصِرة، وأسهمَ في أكثر من عشرة مشرُوعاتٍ بحثيَّةٍ دوليَّةٍ في ميادين مُعالجَة اللُّغات الطَّبيعيَّة. حصلَ على عددٍ من الجوائز في مَيدان تخصُّصه، منها: جائزة (ألكسو ALECSO) للإبداع والابتكار في «المَعلُوماتيَّة والـمُعالجَة الآليَّة للُّغة العربيَّة»، وجائزة راشد بن حُمَيد للعُلُوم والثَّقافة.

الدكتور/ عبد العاطي إبراهيم هوَّاري

عَمل باحثًا زائرا في جامعة جورج واشنطن، في الولايات المُتَّحدة الأمريكيَّة. حصل على درجة الدُّكتوراه في اللسانيات عام ٢٠٠٨م. عملَ في العديد من المشروعات البحثيَّة العربيَّة؛ كما عملَ باحثًا في جامعة كولورادو وجامعة كولومبيا الأمريكيَّة قبلَ أن يتَّجِّهَ للعمل في جامعة جورج واشنطن. نَشَرَ عددًا من الأوراق البحثيَّة في الدِّلالة المُعجميَّة وقضايا المُعجميَّة العَرَبِيَّة والصرف العربي، كما شاركَ في العديد من المؤتمرات الدَّوليَّة داخلَ مصر وخارجها. له عددَ من المُؤَلَّفات العلميَّة المنشورة.

هذه الطبعة **إهداء من المركز**
ولا يسمح بنشرها ورقياً أو تداولها تجارياً

الموارد اللُّغويَّة الحاسوبيَّة

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجُهد مُحرِّرَي الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المركز؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

الأمين العام للمركز

أ. د. محمود إسماعيل صالح

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف: ٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

9 786038 221549