مؤتمر مجمع الملك سلمان العالمي للغة العربية
حوسبة اللغة العربية وإثراء البيانات اللغوية

المُحَوِّلَاتُ التَّوْلِيْدِيَّةُ العَرَبِيَّةُ
بنَظْرَةٍ تِقَنِيَّةٍ رِيَاضِيَّةٍ بَحثِيَّةٍ مُختَصَرَةٍ

كتابة وتقديم:
سَرِيٌّ بنُ تيسير السِّبَاعي

مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language
برنامج تنمية القدرات البشرية
Human Capaility Development Program
رؤية VISION 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

فهرس العرض

- أهداف البحث
- المدونات النصية
- وصف النموذج
- الاستخدام والتطبيقات
- إضافات لَمْ تُذكر
- خاتمة

"ويأبى اللهُ العصمةَ لكتابٍ غيرَ كتَابه ، والمُنْصفُ من اغتفرَ قليلَ خطأِ المرءِ في كثيرٍ من صَوابِهَ"
ابن رجب الحنبلي (ت ٧٩٥ هـ)

أهداف البحث

- الإشارة لفنّ الترجمة ومركزيَّتِه وُمُعْضِلَتِه.
- جمع شتات شوارد ما للنموذج من مباحث.
- ذكرُ أوابدَ بحثية لم تلحَقها العربية بعدُ.
- تقييد خبرات وفوائد عرضت لي.

و[البحثُ] لمحٌ تكفي إشارته ... وليسَ بالهذرِ طُوِّلَت [فِقَرُه][3]

ملاحظة في الترجمة

أحببت إفرادها بشريحة خاصة فقط لأشير أن هناك ثغرًا عظيما ينبغي سدُّه لا فقط بترجمة حرفية بل بمقابلات تحوي دلاليًّا المُعنى المراد فقد رَجعت في بحثي إلى أربعة معاجم متخصصة بمجال الذكاء الإصطناعي [1] ففيها لا تزال ترى كثيرا من الكلمات تُنَقحر (تُتَرجم صوتيا) [2] مع أن لها تعريبات مقبولةً تشير إلى معانيها، واعلم أن الترجمة الحرفية لا تساعد القارئ ولا الطالب على تصوير المعنى في ذهنه بل تربطه بلغة أخرى بحروف لغته. وأُلمِعُ بالإشارة إلى أن هناك مئات المعاجم العربية في مختلف العلوم نغفل عنها تُفيدنا في ربط العلوم بعضَها ببعضٍ ولا سِيَّما ألا علمَ غيرَ بينيٍّ في زماننا، ولكنَّ هذا يحتاج جهدًا جهيدا وسعيا شديدًا وتفرغا تامًّا واستحضارًا عامًّا بجرد المطولات وجمع الفوائد والنكات ولمِّ شعثِ المتناثر بجهودٍ متتابعة تتظاهر وتتعاسسُ

وَقُل هَل فَشا في الأَرضِ غَيرُ لِسانِهِم — لِسانٌ فُشُوَّ الضَوءِ وَاليَومُ شامِسُ [4]

المدونات النصية

المدونات النصية

أشهرها[6]

هذه بعضٌ من أكبر المدونات النصية الخام المنشورة:

- مدونة CaltureX
- مدونة ArabicText2022
- 101 Billion Arabic Dataset * مدونة

تمثيلها

أشهر تمثيل للنصوص المكتوبة يكون باستعمال خوارزمية BPE

Byte Pair Encoding
التزويج البايتي

وعليها بُنيَت الباقيات، مثلَ SentencePiece

وينبغي دراسة المقسمات بما يتناسب مع العربية

المدونات النصية

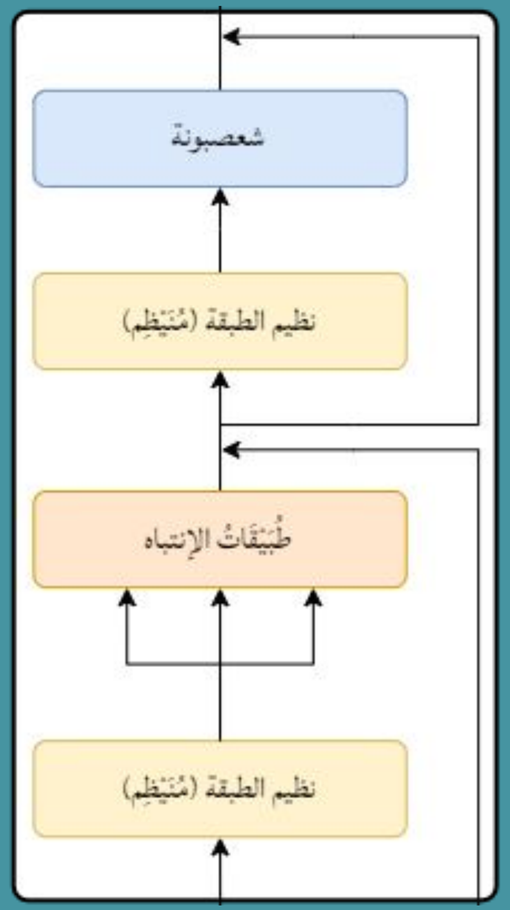
مشاكل المدونات

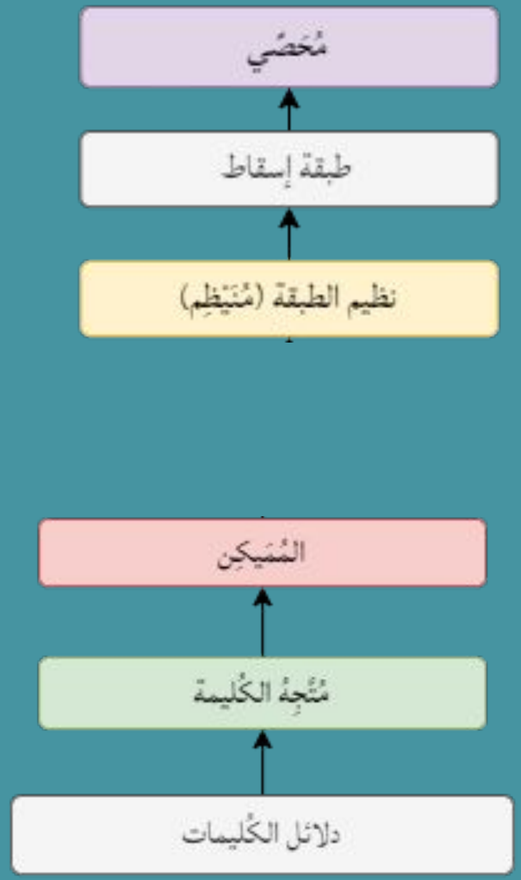
- قلة النصوص
(أكبر الموجود مدونة سدايا 540 مليار جزء لغوي ونصفُه مترجم) والإنجليزي فمن المنشور له 15 ترليار جُزءٍ لغوي
- المحتوى غالبه من الشبكة وليس بذي جَودةٍ. وكمية الكتب قليلة (من منشور وتجربة)

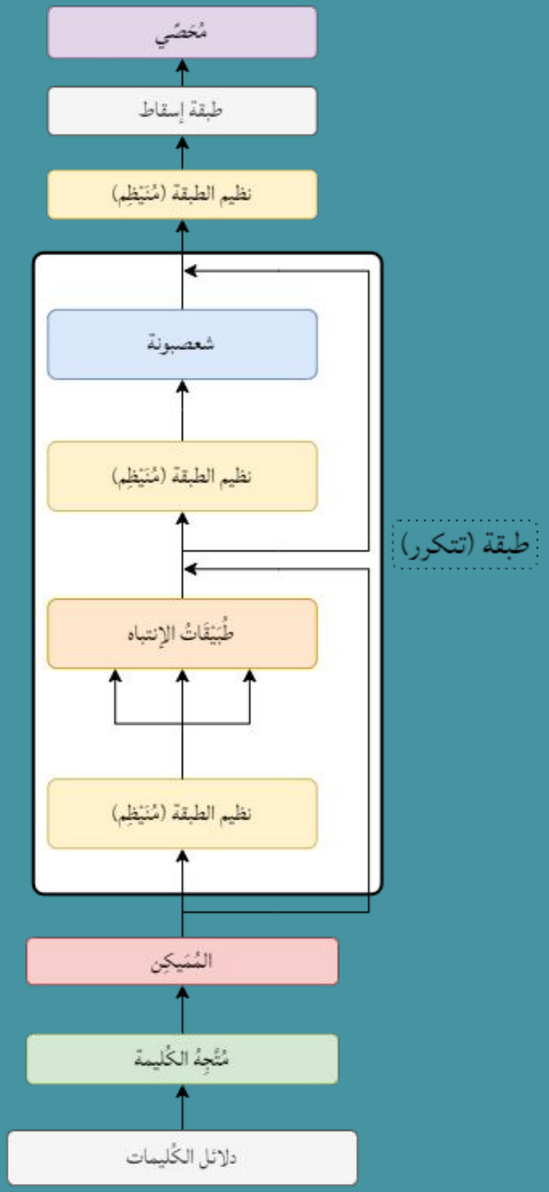
مُقترحات

- تكريس الجهود في تحسين الترجمة يدويًّا ثم آليًّا.
- تطوير أساليب استخراج النصوص من المصورات
- إرجاءُ تطوير النماذج والتركيز على رقمنة النصوص، والترجمة والفهرسة.

نموذج المحوِّل


شعصبونة
نظيم الطبقة (مُنَيْظِم)
طُبَيْقَاتُ الإنتباه
نظيم الطبقة (مُنَيْظِم)


مُحَصِّي
طبقة إسقاط
نظيم الطبقة (مُنَيْظِم)
المُسَيكِن
مُتَّجِهُ الكُليمة
دلائل الكُليمات


مُحَصِّي
طبقة إسقاط
نظيم الطبقة (مُنَيْظِم)
شعصبونة
نظيم الطبقة (مُنَيْظِم)
طُبَيْقَاتُ الإنتباه
نظيم الطبقة (مُنَيْظِم)
طبقة (تتكرر)
المُسَيكِن
مُتَّجِهُ الكُليمة
دلائل الكُليمات

$$\mathcal{L}(\hat{y}, y) = \text{CrossEntropy}(\hat{y}, y)$$

$$Z_2 = F_2 W_2 \quad Z_2 \in \mathbb{R}^{T \times d_{model}}$$

$$Z_3 = Z_1 + Z_2 \quad Z_3 \in \mathbb{R}^{T \times d_{model}}$$

$$Y = \text{Softmax}(Z_3 W_Y) \quad Y \in \mathbb{R}^{T \times d_{vocab}}$$

$$F_1 = \hat{Z}_1 W_1 \quad F_1 \in \mathbb{R}^{T \times d_{ff}}$$

$$F_2 = f(F_1)$$

$$Z_1 = Z_0 + \text{MultiHead}(Q, K, V)$$

$$\hat{Z}_1 = \text{LayerNorm}(Z_1) \quad \hat{Z}_1 \in \mathbb{R}^{T \times d_{model}}$$

$$Q = \hat{Z}_0 W_Q \quad K = \hat{Z}_0 W_K \quad V = \hat{Z}_0 W_V$$

$$Q, K, V \in \mathbb{R}^{T \times d_k}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) V \quad M \in \mathbb{R}^{T \times T}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O$$

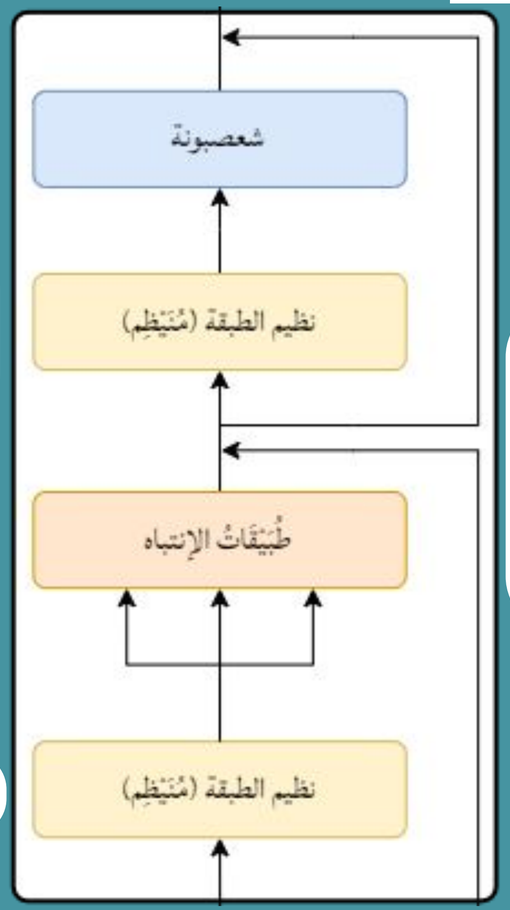
$$W_O \in \mathbb{R}^{(h \times d_k) \times d_{model}}$$

$$\hat{Z}_0 = \text{LayerNorm}(Z_0) \quad \hat{Z}_0 \in \mathbb{R}^{T \times d_{model}}$$

$$E = \text{Embedding}(X) \quad E \in \mathbb{R}^{T \times d_{model}}$$

$$PE = \text{PositionalEncoding}(T, d_{model})$$

$$Z_0 = E + PE \quad Z_0 \in \mathbb{R}^{T \times d_{model}}$$



ملاحظات المحولات

- انتشرَ كثيرًا الآن استخدام هيكلية LLAMA وفروعها لأنها مفتوحة المصدر وتتابع العَمَلُ واستعمالها.
- غالب العمل محاولاتٌ وتجارب في التدريب تحت مظلة المستنتجات التجريبية.
- اعتماد هذه الهيكلية جعلَ الالتفاتَ لمُخلتف الجديد قليلًا مثل "مَمْبَة Mamba" أو "رَوْكَفَ RWKV".
- نحنُ افترضنا توزيعًا احصائيا وسرنا عليه (باستخدام "المُحَصِّي softmax")

  فما بقيَ من التعلم الإحصائي والفكر اللساني؟
- النموذج نتاج مجموع حلول بعض المشاكل السابقة (مثل RNN) والتقدمات التقنية (القوة الحوسبية) فاستشرا.

- محاولة تفسير أداء النموذج متعثرة بتعثر سابق في تعليل سلوك نماذج الصُوَر، ولكنَّنا بدأنا فهم حَدَّ قدراتها.
- كثير من النماذج العربية تَصْدُرُ ثمَّ تُترك بعكس التوجه الربحِي والذي مآلهُ إنتاج تجاري.
- هذه النماذج موسوعات متكلِّمَة -مانِنْجُ Manning- ضاغطة لا مُتَفَهِّمَة.
- بعدَ انفقاع فقاعة "الذكاء العام"، ستبقى هذه كأدوات، وسنبدأ حقبة جديدة كالعادة.
- مَن لَّا يَملِكُ 100 مليون فلا يُكَلِّفَنَّ نفسه مالا تطيق.

المعيرة والتطبيقات

المعيرة (الضبط الدقيق)

بأن تُعَوِّدَ النموذج على أسلوب المحادثة (أو غيره) بإمراره على بيانات مُهَيكلة (ذات هيئة تتابعية تُشبِه الدردشة).

وبعدها مرحلة بَعدَ المعيرة يُقَيَّد فيها جواب النموذج بحسبِ ثقافة الجهة المُنتجة وهذه يُستَعمل فيها عادةً علم "التعلم المعزز (الجزائي)" وله أساليب كثيرة وهذا فَنٌّ لوحده.

وفي كلا المرحلتين تبرز تحديات المشاكل الثقافية (ويأتي غالبها من المترجم)، والصحة المعرفية (وغالبها من جودة النصوص)

لعل سائلا يسأَل "وكيف يُجيبُ على ما لم يَرَهُ في مرحلة المَعيرة؟" فنقول 'أنه يستحضره مما تدرب عليه في المرحلة الأولى' فإن سألَ "وكيف يُدرك ذلك " لقُلنا 'أننا لا ندري للآن تفاصيل ذلك 🙂".

تطبيقات

لا يَخفى فشُوُّ انتشار استخدام مُنتجات هذه النماذج وأشهرها ChatGPT وأصول الاستخدام ترجع **للتطبيقات اللغوية الأساسية** (الاستجابة للأمر) ولما استجدَّ من استخدامات في كتابة البرمجيات وترتيب البيانات وهيكلتِهَا، وأخيرا توليد النصوص، كُلُّ ذلك **بنَصٍّ (أمرٍ) معروضٍ** عرْضًا مُفصَّلًا أبلج الملامح دقيقًا سافرَ اللوائح.

ولكن من المهم الإشارة لمبدأ **التعويل والاتِّكال** وهو ألا يُنظر لهذه المُنتجات كبدائل بل أدوات ولا غايات بل مُساعدات فلا يُتَّكَؤُ عليها بل يُستفاد منها، **وهنا تظهر أفكار بحوث كثيرة تقنية ونفسية وتعليمية.**

إضافات وزوائد

إضافات وزوائد

"إني رأيت أنه لا يكتب إنسان كتابًا في يومه إلا قال في غده: لو غُيِّر هذا لكان أحسن، ولو زيد كذا لكان يُستحسن، ولو قُدِّم هذا لكان أفضل، ولو ترك هذا لكان أجمل. وهذا من أعَظم العبر، وهو دليل على استيلاء النقص على جملة البشر."

العماد الأصفهاني (ت ٥٩٧ ه)

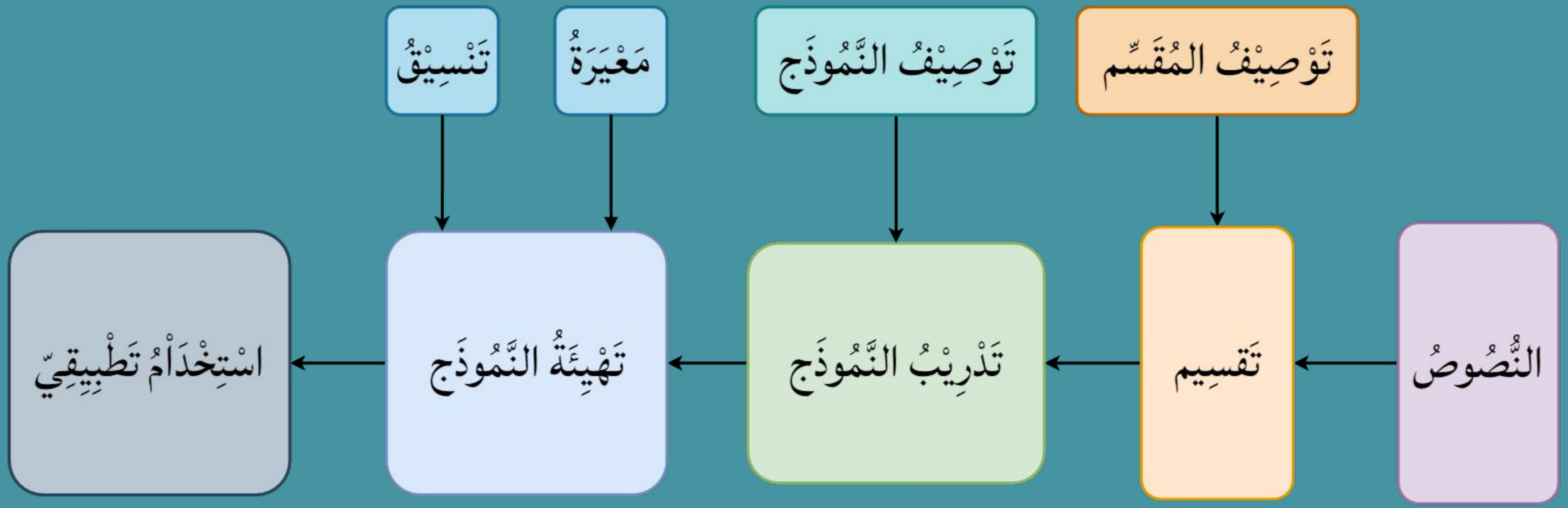
فعلى سبيل التمثيل لا الإحصاء.

- تحديثات هندسة الأوامر في سلسلة الأفكار (تأمل الأوامر)
- اختراق النماذج وَوُهُنها
- تطبيقات استرجاع المعلومات
- استرداد النصوص المُتَدَرَّب عليها من النموذج
- تقييم النماذج وأنواعها
- جودة المدونات النصية وطُرقُها
- تقطير النماذج واستخداماته.

ملخص

ملخص

رَأَيْتُ النَّاسَ غَالِبَهُمْ رُسُومِيْ... ولَيْسَ بِقَارِئٍ يَفْرِيْ السُّطُورَا



وكُلُّ جُزءٍ مبحثٌ يفنى فيه الشادي والباحث

سؤال العرض

مَقصَدُنا السؤال والنقاشُ ... لا كثرةُ الكلام والرَّقاشُ

مراجع الشرائح

1. (1) معجم سدايا (2) ومعجم علاء طعيمة ،(3) ومعجم الذكاء الإصطناعي لمكتبه في الإمارات (4) قاموس موقع الذكاء الاصطناعي باللغة العربية.

   وهناك مراجع أخرى: كمعجم رياضيات دمشق والقاهرة وموقع المعاني والمورد الكبير والحديث والعادي وغيرها.
2. أي تُنقل حرفيا انظر مقالة النقحرة على موقع ويكيبيديا: https://ar.wikipedia.org
3. مقتبس من بيت شعر (رقم 16) في قصيدة (رقم 68) أولها "لا الدهر مستفذٌ ولا عجبه ... تسومنا الخسفَ كُلَّهُ نُوَبُه" من ديوان البحتري (٢٠٧/١-٢١٠)، تحقيق حسن كامل الصيرفي، دار المعارف، (وأشيد بموقع الشنكبوتية لتسهيل الوصول)
4. من قصيدة للزمخشري في مدح العرب أولها "أيا عرصات الحيّ أين الأوانس ... رحلت وحَلَّتك الظباء الكوانس" في ديوانه صـ298 قصيدة رقم (4) والبيت رقمه (33). طبعة دار صادر
5. ورقة علام المنشورة: ALLaM: Large Language Models for Arabic and English
6. منشورة كلها على موقع HuggingFace

لمراجعة الشرائح

مؤتمر مجمع الملك سلمان العالمي للغة العربية
حوسبة اللغة العربية وإثراء البيانات اللغوية

شكراً لكم

مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language

برنامج تنمية القدرات البشرية
Human Capaility Development Program

رؤية VISION 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA