

Language Modelling from Pixels

Desmond Elliott

Language and Multimodal Processing Group
Department of Computer Science
University of Copenhagen



Warning: The final part of the talk contains dataset samples that are racist in nature.

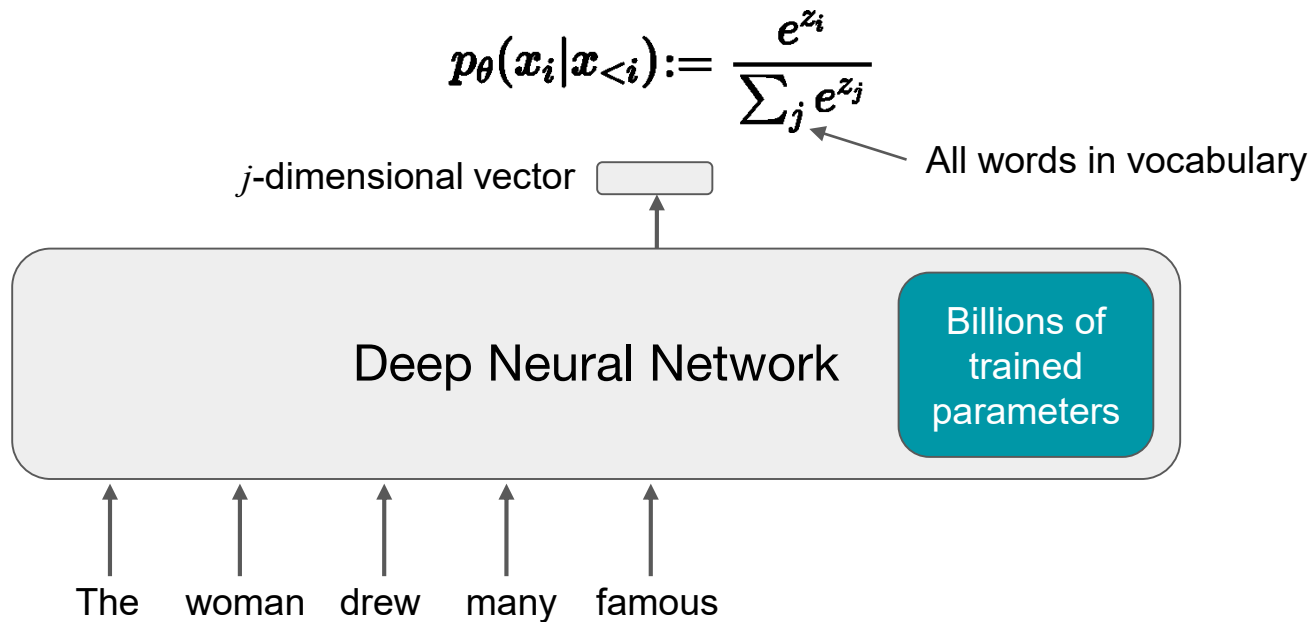
What is Language Modelling?

- A language model is a statistical model of text and can be used to estimate the probability of a string \mathbf{x} consisting of T tokens.



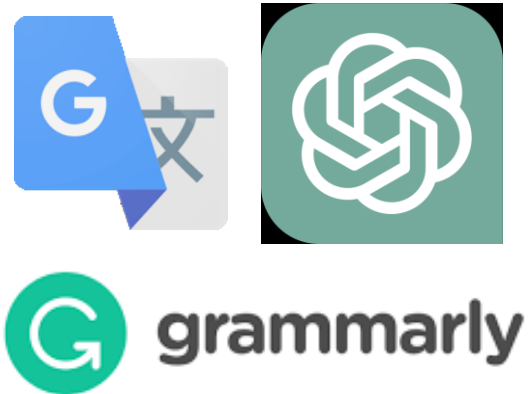
Tractable with assumptions,
e.g. bi-gram model

Modern Language Models

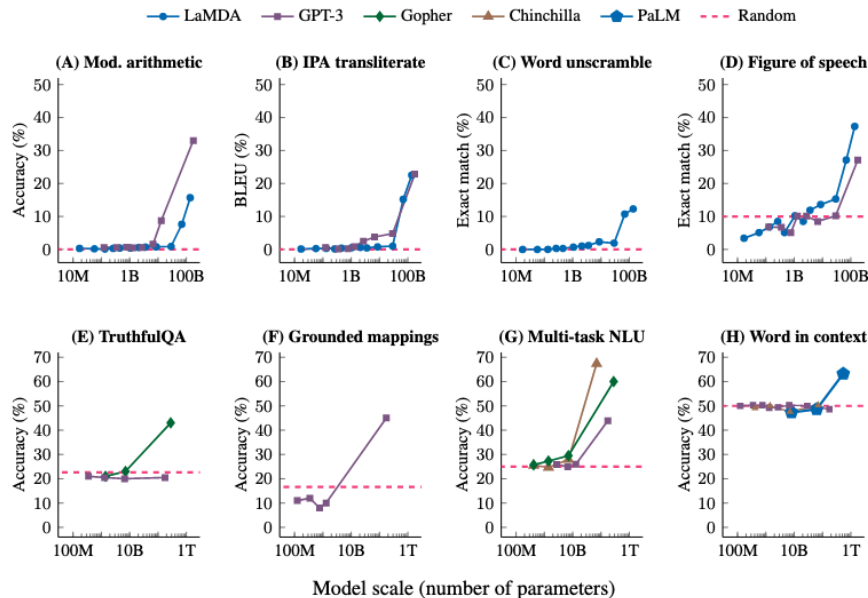
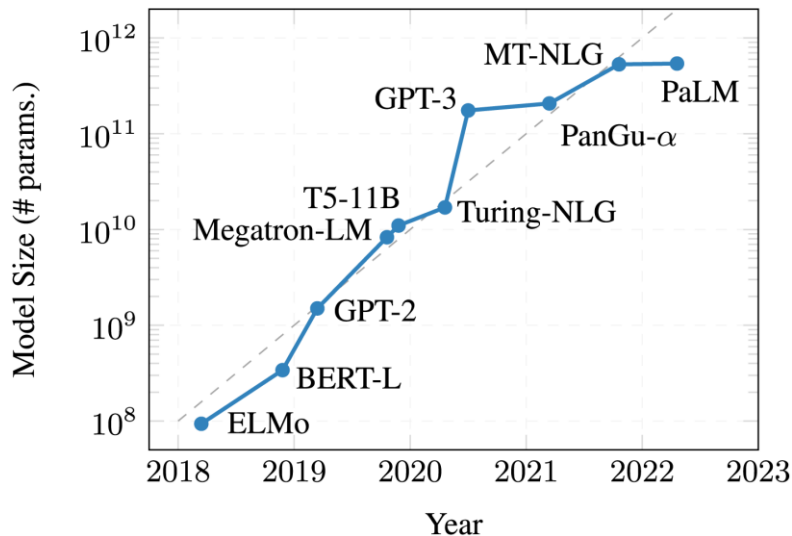


Okay but why is it useful?

- Spelling and grammar checking
- Machine translation
- Web search
- Text prediction
- Topic modelling
- Chatbots

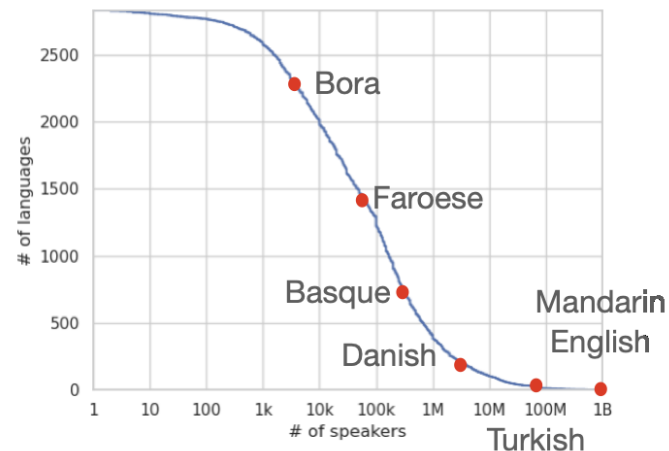


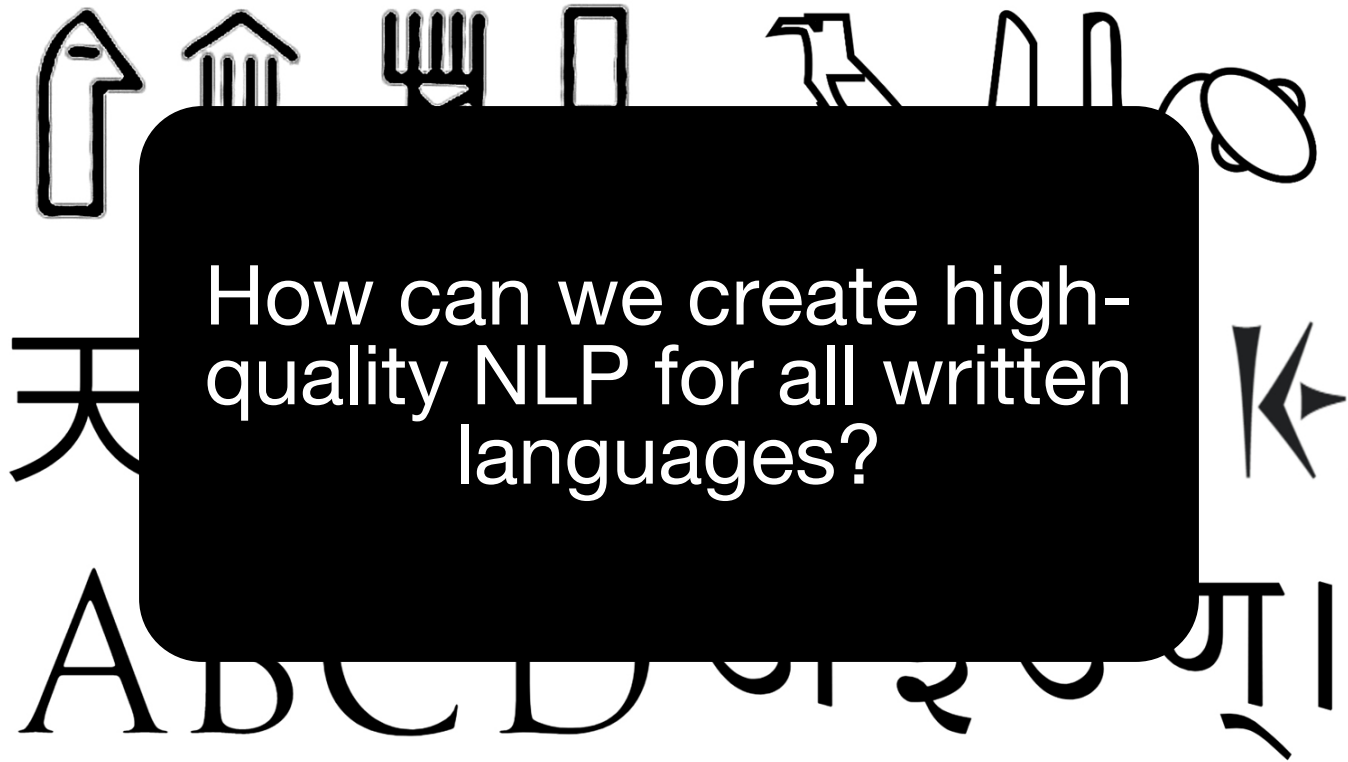
Current Status: Scale



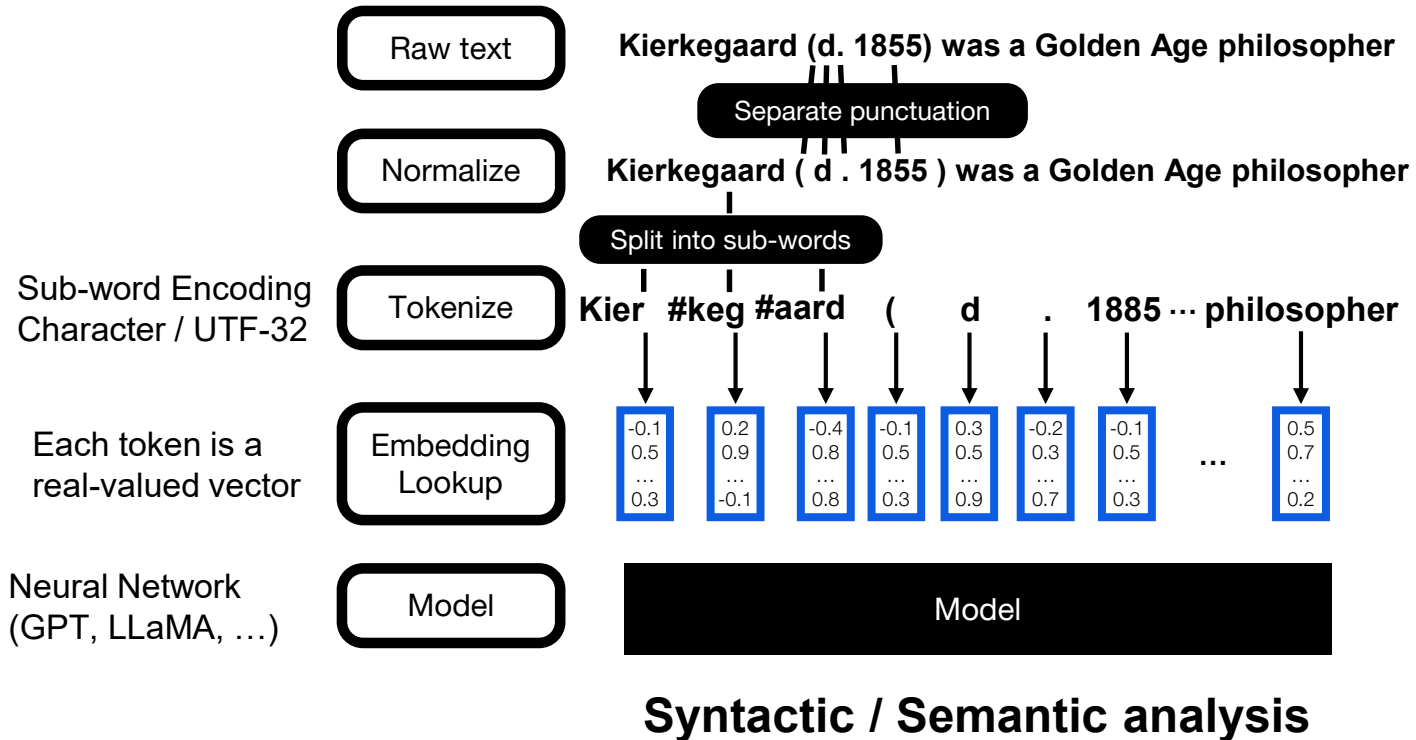
NLP for **All** Written Languages

- There are 3,000 written languages
 - 400 with >1M speakers
- NLP usually covers 100 languages
 - Technological exclusion for billions

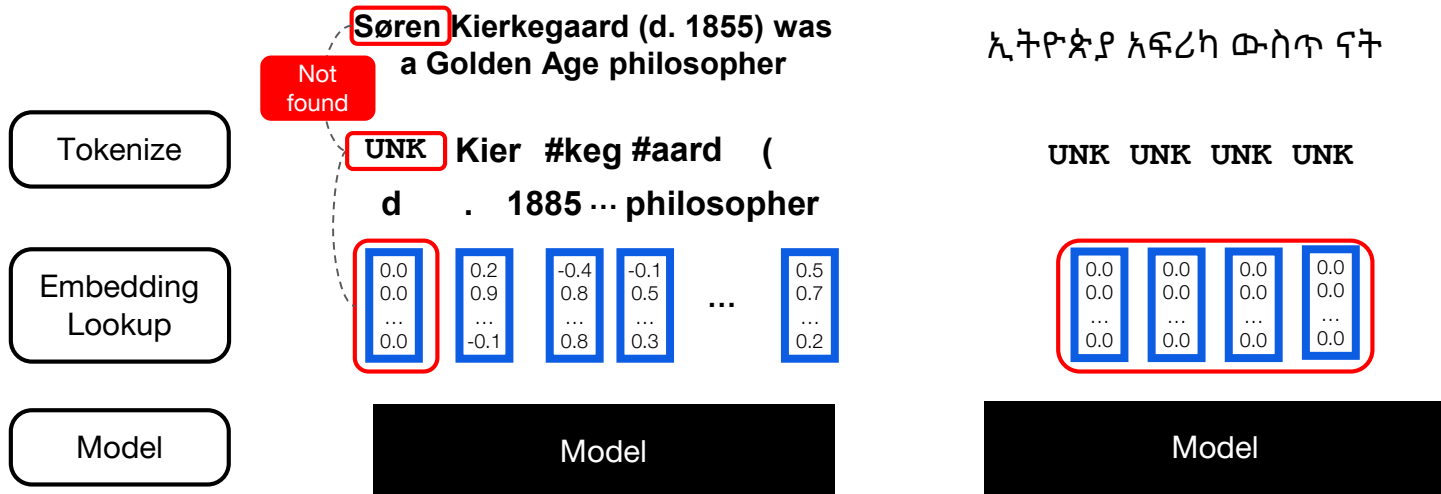




NLP is a pipeline ...



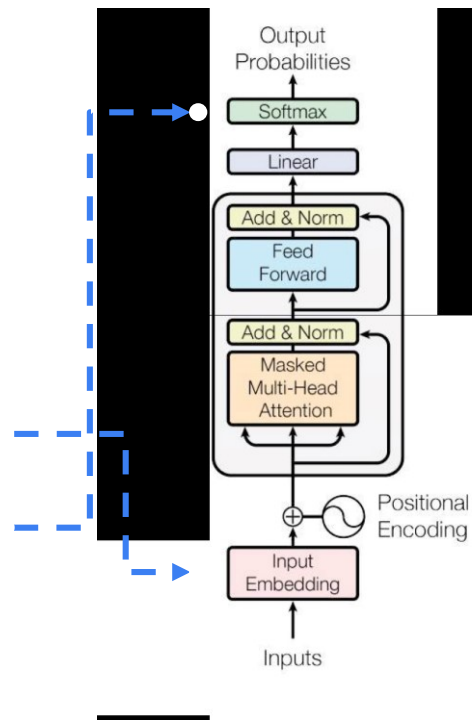
... that is easily broken



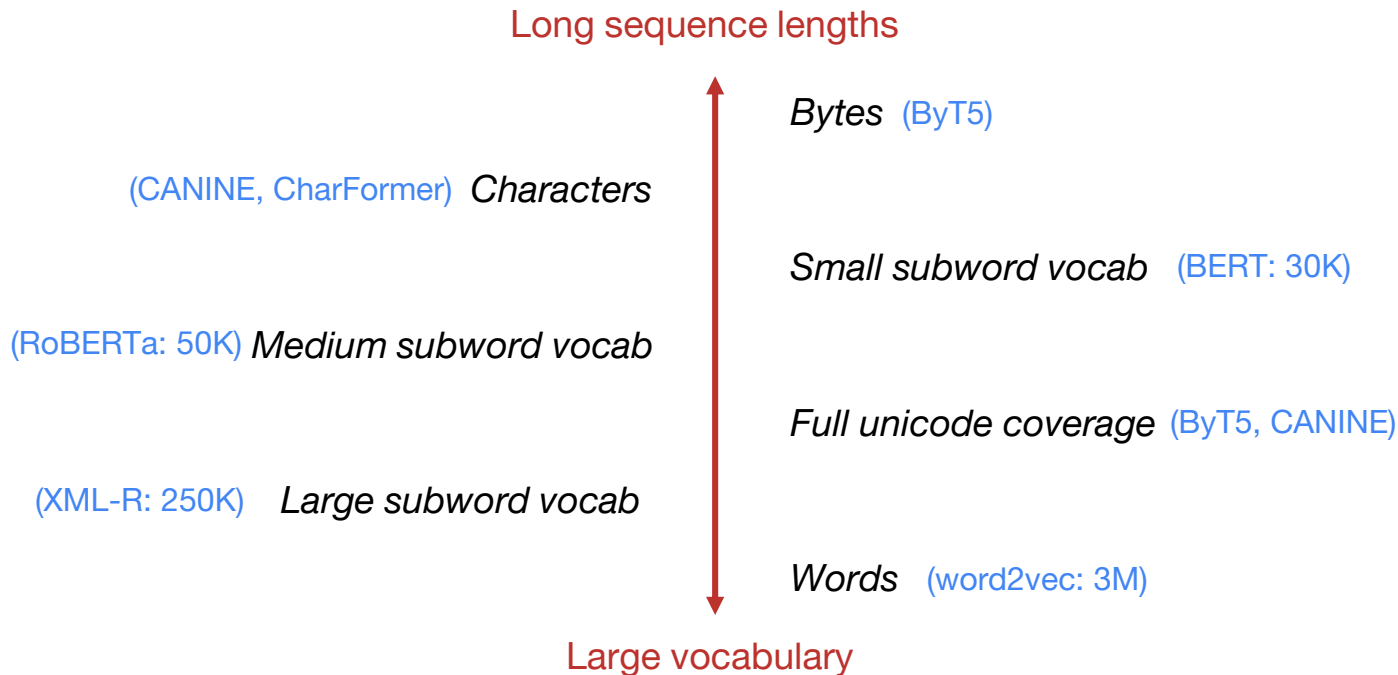
This issue disproportionately affects low-resource languages

The Vocabulary Bottleneck

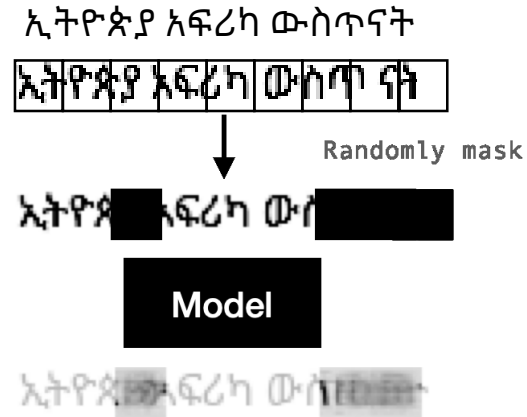
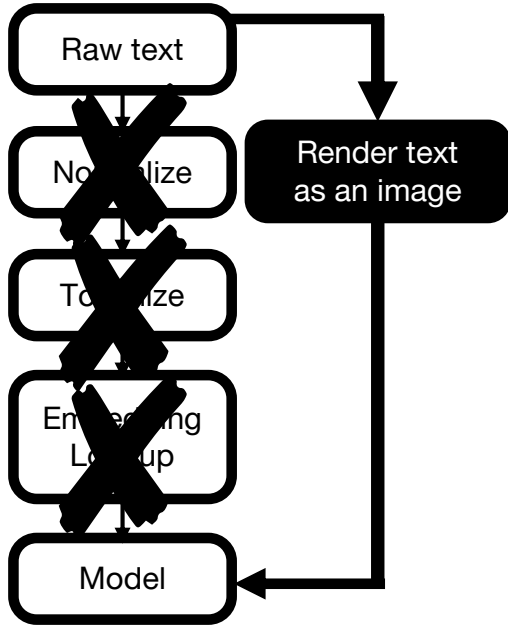
- NLP is an **open vocabulary problem** and the ability of a model is determined by its vocabulary:
 1. tokens, characters, sub-words, etc.
- This creates a bottleneck in two places:
 1. *Representational bottleneck* in the Embedding layer
 2. *Computational bottleneck* in the Output layer



Where's the sweet spot?



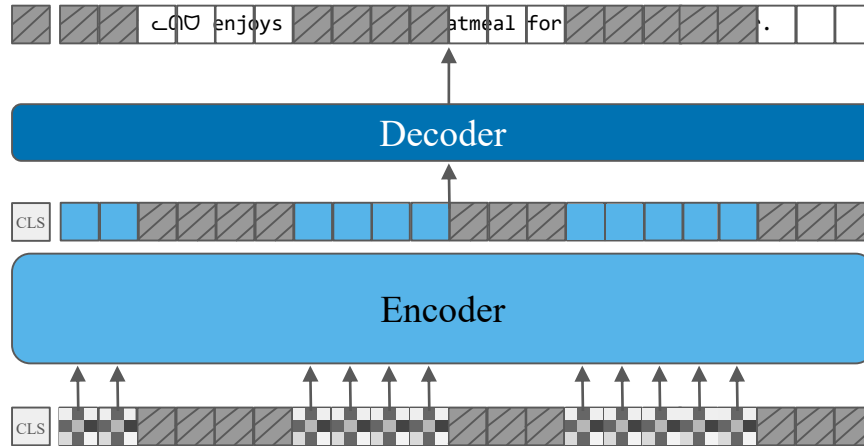
Treat language as vision



Overview

1. Pixel Language Model
2. Text Rendering Matters
3. Historical Document Processing

The Model



8 Layers

12 Layers

- 3 CLS Embedding & Span Mask m patches
- 2 Projection + Position Embedding

- 1 Render Text as Image

My cat ຈຸດ enjoys eating warm oatmeal for lunch and dinner.

16pixel x 16pixel patch

Google Noto Fonts

PyGame / PangoCairo

Flexible Text Renderer

My cat 🐱 loves pancakes 🥞 and my duck 🦆 loves grapes 🍇. ■■■

- Emoji

- Left-to-right text: 牠們常在晚間活動，但並不表示他們是夜行性動物。 ■■■ تنشط القطط في الخلاء ليلا ونهارا على الرغم من أنها تميل إلى أن تكون أكثر نشاطا بقليل في الليل. ■■■

- Wcwidth: ደመት በአሁኑ ጊዜ ከሁሉም እንስሳ በላይ በቤት እንስሳነቱ ተፈላጊነትን ያላት ናት ። ■■■

A new type of generative model

Penguins are designed to be streamlined and hydrodynamic, so **having thin legs** would add expanding. Having **short legs** with **wedged feet** to act like **rubbers**, helps to give them that **top do-like figure** **didn't** compare bird anatomy with humans, we would see **something** **is** peculiar. By taking a look at the side-by-side **image** in Figure 1, you can see how their leg **bones** **are** to ours. What most people mistake for **knees** are actually the **anatomies** of birds. This **gives a conclusion** that bird knees bend opposite of ours. The knees are actually tucked up inside the **bones** of the **bird**. So how does this look inside of a penguin? In the **images** below, you can see boxes surrounding the penguins' knees.



Penguins are designed to be streamlined and hydrodynamic, so **having long legs** would add expanding. Having **short legs** with **wedged feet** to act like **rubbers**, helps to give them that **top do-like figure**. If we compare bird anatomy with humans, we would see **something** **is** peculiar. By taking a look at the side-by-side **image** in Figure 1, you can see how their leg **bones** **are** to ours. What most people mistake for **knees** are actually the **anatomies** of birds. This **gives the illusion** that bird knees bend opposite of ours. The knees are actually tucked up inside the **bones** of the **bird**. So how does this look inside of a penguin? In the **images** below, you can see boxes surrounding the penguins' knees.

100K steps

500K steps

1M steps

Pretraining

- **English Dataset:** English Wikipedia and Books Corpus
- **Masking:** 25% Span Masking
- **Maximum sequence length:** 529 patches (16x8464 pixels)
- **Compute:** 8 x 40GB A100 GPUs for 8 days
- **Parameters:** 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The Great Wall of China (traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: Wànlǐ Chángchéng)

Downstream Tasks

- **Datasets:** Universal Dependencies, MasakhaNER, GLUE, Zeroé

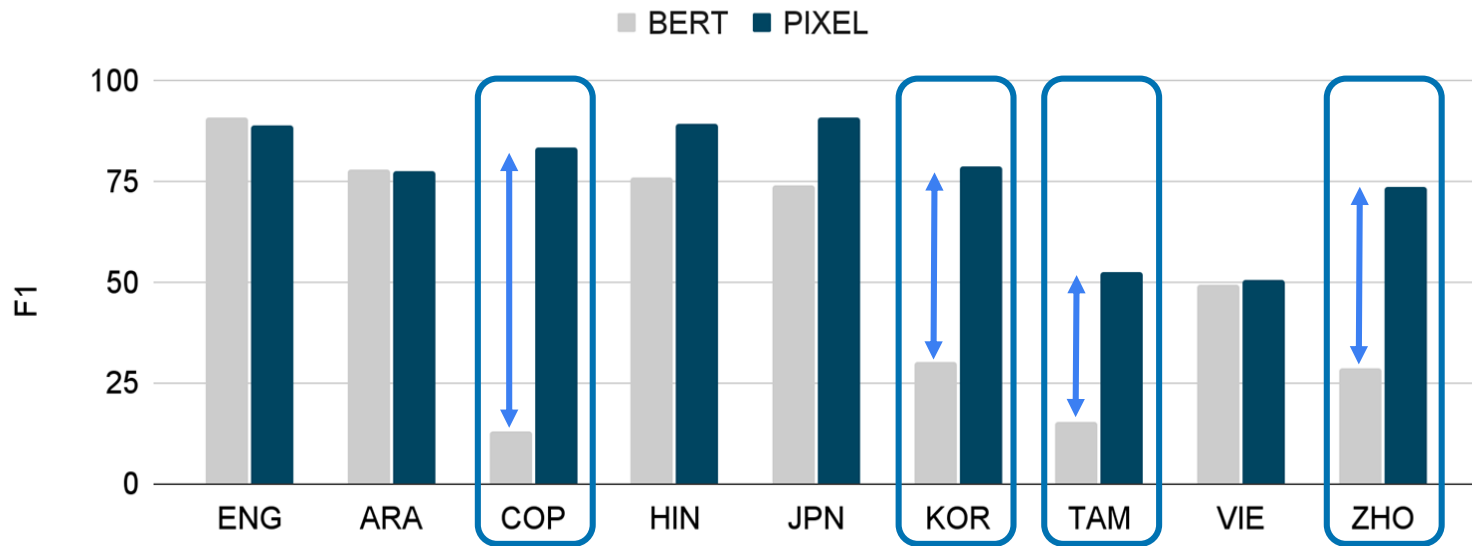
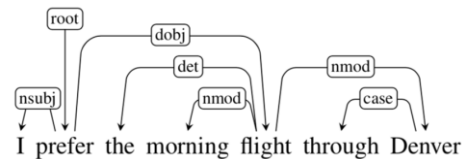
- **Models:**

	Parameters	Pretraining Data
PIXEL _{BASE}	86M	English Wikipedia + Bookcorpus
BERT _{BASE}	110M	—
CANINE-C	127M	104-languages from Wikipedia

Similar pretraining setup

Tries to solve the same problem using UTF-32

Dependency Parsing Results



BERT UNK

0%

1%

94%

33%

46%

85%

82%

5%

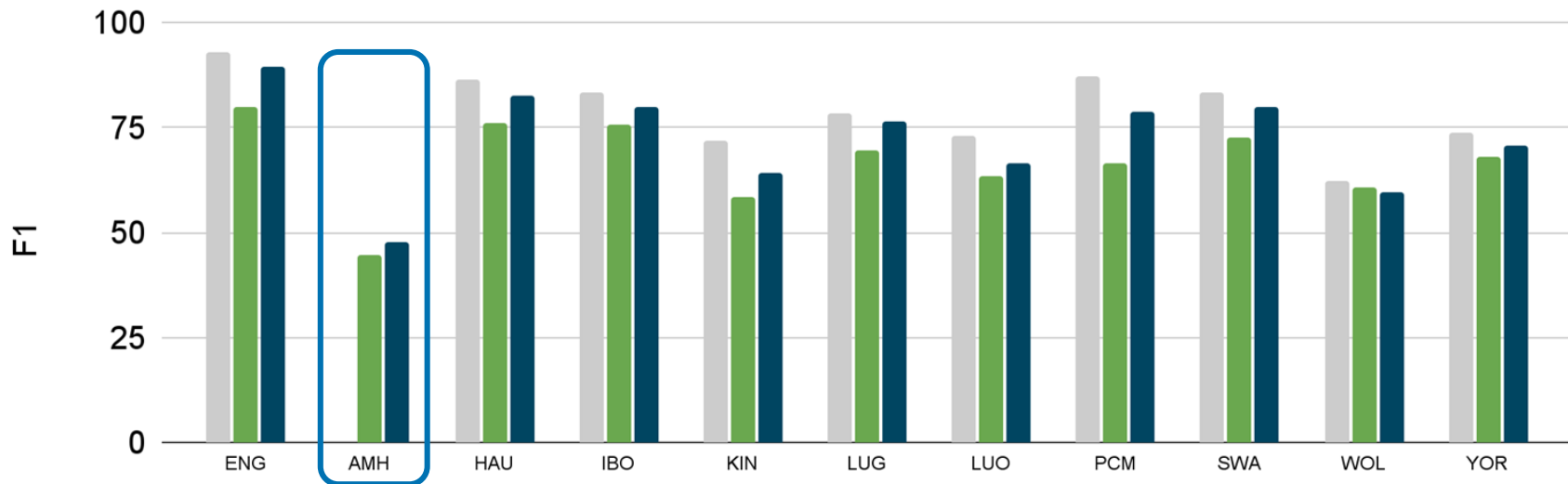
73%

PIXEL vastly outperforms BERT on unseen scripts

Named Entity Recognition in African Languages

Emir of Kano turban Zhang wey don spend 18 years for Nigeria

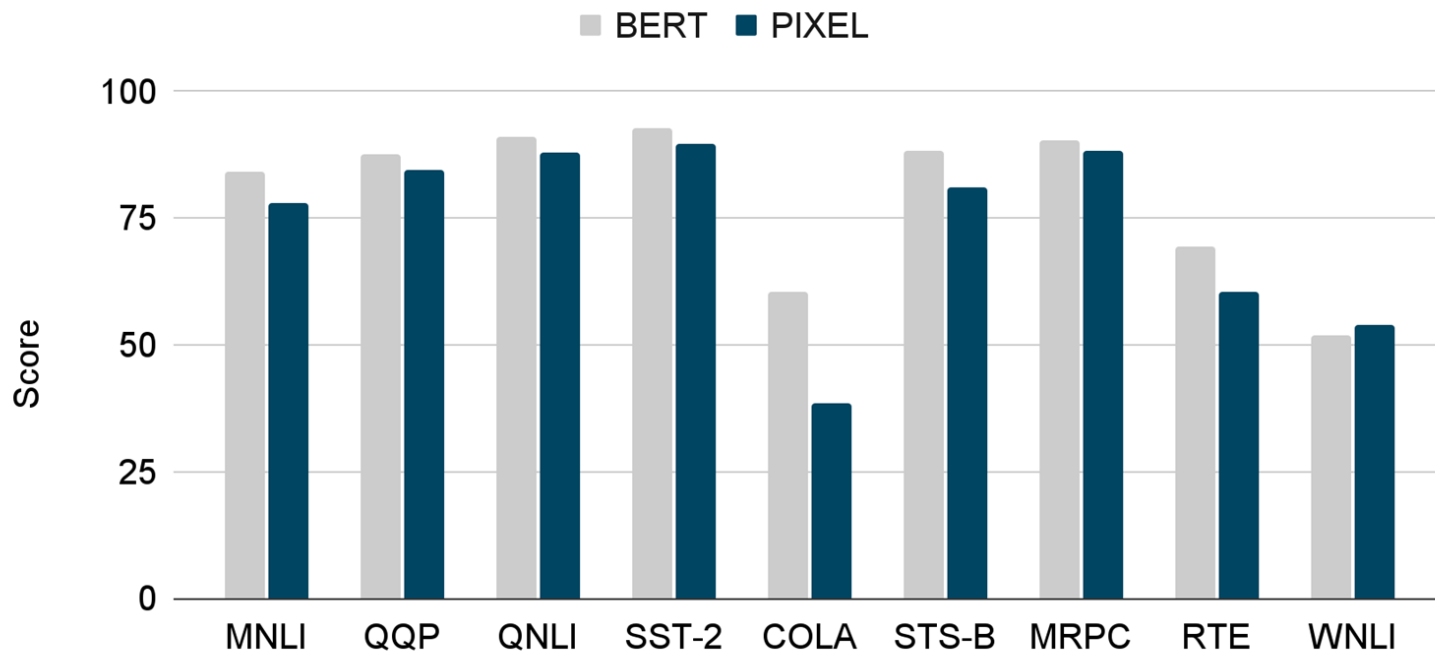
■ BERT ■ CANINE ■ PIXEL



PIXEL outperforms BERT on the non-Latin script

PIXEL outperforms the multilingually pretrained CANINE-C

GLUE: Sentence-level Understanding



BERT outperforms PIXEL on English sentence-level tasks

2. Text Rendering Matters

Text Rendering Matters

- Our original text renderer produces many nearly-identical patches
 - This is representation- and compute-wasteful

the the the the the the the the the

Can we do better?

Alternative Text Renderers

(a) Continuous rendering (CONTINUOUS):

I m u s t b e g r o w i n g s m a l l a g a i n . ■

(b) Structured rendering (BIGRAMS):

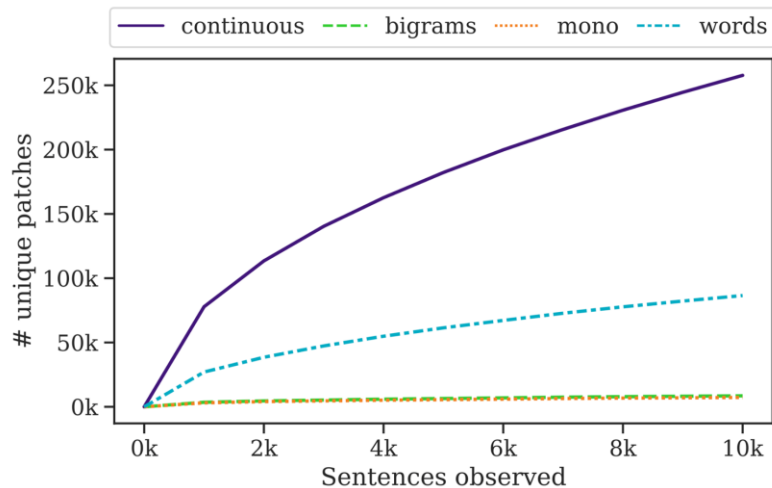
I m u s t b e g r o w i n g s m a l l a g a i n . ■

(c) Structured rendering (MONO):

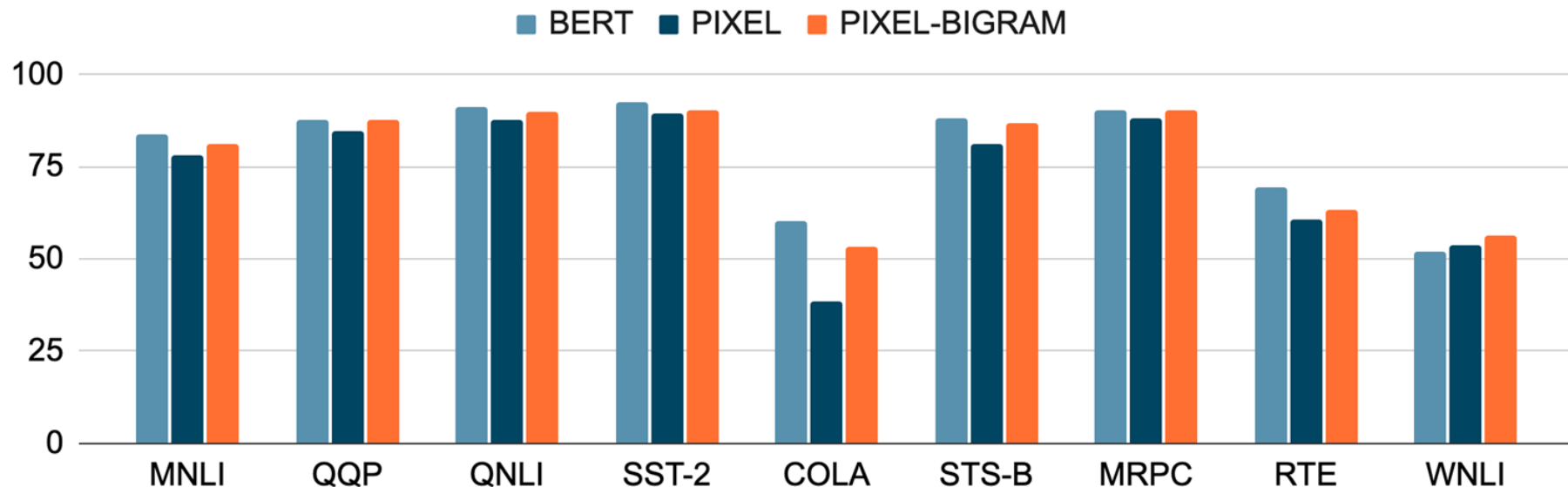
I m u s t b e g r o w i n g s m a l l a g a i n . ■

(d) Structured rendering (WORDS):

I m u s t b e g r o w i n g s m a l l a g a i n . ■



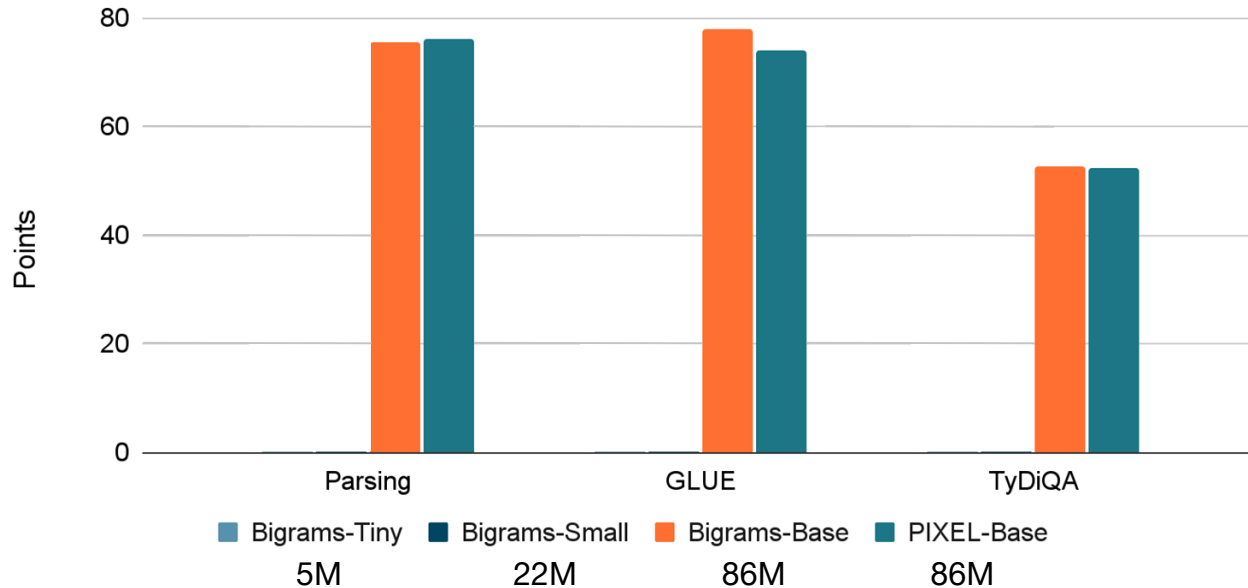
GLUE (revisited)



Bigram text rendering produces better models

Scaling Down

- Better text rendering can create effective models at smaller scales



3. Historical Document Processing

Warning: This part of the talk contains dataset samples that are racist in nature.

Historical Document Processing

- Worldwide efforts to digitize historic documents (Groesen 2015)
- Typical pipeline for enabling access is:
 - a. Scan documents into high-quality digital formats
 - b. Perform OCR on those documents (one-off process)
 - c. Search through documents using OCR annotations

What if we could do this without OCR?

Caribbean Newspapers, 1718-1876

- Collaboration with researchers that are interested in tracking newspapers notices about escaped slaves

- What was the given name?
- What reward was offered?
- Who was the contact person?

- Dataset of 1.65M scanned pages



PHD: PIXEL for Historical Documents

- Historical document-aware Pretraining
 - Mixture of scanned newspapers and synthetic newspaper-like text generated from Wikipedia and Bookcorpus datasets
 - All input data is scaled to 368x368 and split into 16x16 patches

sionally blogs such as Arcade, a humanities site published by Stanford University. From 2012 to 2016, he hosted a radio show webcast by Alanna Heiss's Clocktower Productions. In autumn 2020, an article he wrote for The Creative Independent was widely disseminated on the internet. Called 19 things I'd tell people contemplating starting a record label (after running one for 19 years) it was a mix of advice, warnings, and personal history gleaned from almost two decades of operating Brassland. It was followed by an appearance on the Third Story podcast.

Sickman's war service took him to Tokyo during the occupation of Japan where he served as one of the "Monuments Men" under General Douglas MacArthur's

terminated by the FC England club in 1981 in order for The Championships, Wimbledon to be held. Since then the club has been nomadic, moving to Osterley and Greenford before settling in Acton and playing their matches at Wasps FC's Tuford Avenue Sports Ground. By 2012, the club had downsized to running only one team.

A number of players for the New Zealand national rugby union team have played for London New Zealand including Doug Rollerson, Terry Morrison and Paul Sapsford. In recognition of their history, the club have been granted privileges from both the Rugby Football Union and the New Zealand Rugby. They are the only rugby team aside of New Zealand national representative teams that were the silver fern as their crest and the RFU exempted them from the overseas player quotas prior to their abolition. The club have also taken part in a number of New Zealand government

aving been estranged from her father's family for most of her life, Andrea is intrigued. But what exactly is the Bancroft's involvement with "Genesis," a mysterious person working to destabilize the geopolitical balance at the risk of millions of lives? In a series of devastating coincidences, Andrea and Belknap come together and must form an uneasy alliance if they are to uncover the truth behind "Genesis"—before it is too late.

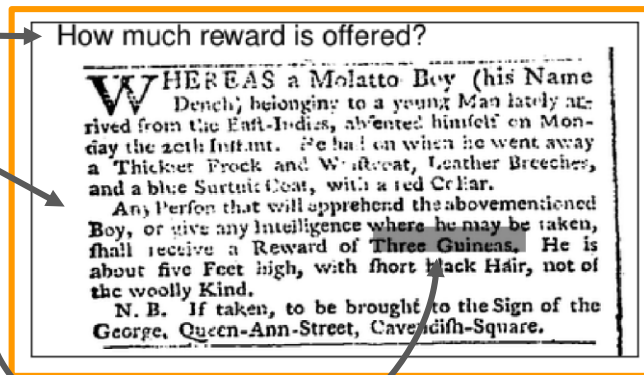
Girls' BMX was part of the cycling at the 2010 Summer Youth Olympics program. The event consisted of a seeding round, then elimination rounds where after three races the top 4

swimmers have so far achieved qualifying standards in the following events (up to a maximum of 2 swimmers in each event at the Olympic Qualifying Time (OQT), and potentially 1 at the Olympic Selection Time (OST)):

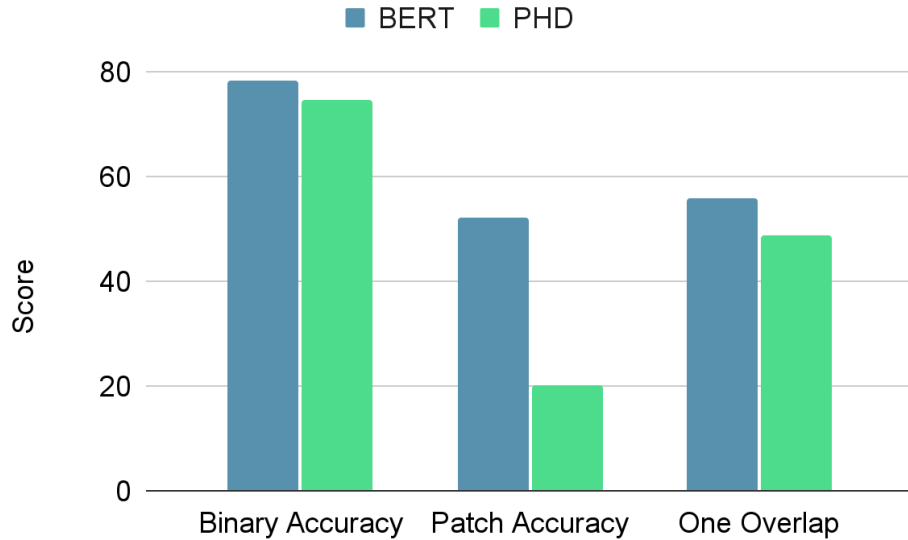
Venezuela has entered one athlete into the table tennis competition at the Games. Gremlins Arvelo secured the Olympic spot in the women's singles by virtue of her top six finish at the 2016 Latin American Qualification Tournament in Santiago, Chile.

Visual Question Answering in Newspapers

- Frame this as a Visual Question Answering Task
 - Render the question
 - Render the clipping on a canvas
 - Annotate context of answer
- Train the model to predict the label of the answer



Results



What other rewards were offered?
R U N A W A Y,
From the Ship BRITANNIA, Capt. Scott,
Commander, on Friday the 25th Instant,
TWO Negro Men, the one named
LEWIS, near Six Feet high, and two
Holes in his Ears; the other about Five Feet
Six Inches high, he has two or three Particular
Sears between his Eyebrows, and his Teeth are
filed down like a Saw between every Tooth. If
any Body will bring them to Mess. MUER and
CLANDEK, Merchants, in Nicholas Lane,
shall be **handlomely rewarded**.

Who is the contact person for the ad?
Last week run away from his Master **J. Bromley**, Esq,
of Bookham in Surrey, his Negro Man **Prince**, alias
Harry Johnson, aged about 35 Years, tall of stature, some-
times wears a Perriwig, speaks English well, in a blue
Livery with Erals Buttons, and has taken with him feve-
ral of his Masters Goods. Whoever secures him, and
gives notice to his Master aforesaid, or to Mr. **Richard**
Sheppard in **Lorbury**, London, shall have a Guinea Re-
ward.

Surprisingly good performance compared to a model
trained on manually transcribed text

Conclusions

- PIXEL is a new type of language model that tackles the open vocabulary problem using visually rendered text
 1. This enables high-quality transfer to different scripts
 - New, unseen languages
 - Different fonts in existing languages
 2. Compact models with as few as 5M parameters
 3. Natural interface to scanned documents

Open Questions

Thanks



N. Borenstein



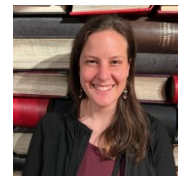
P. Rust



I. Augenstein



E. Bugliarello



E. Salesky



M. de Lhoneux



Models



Code