Recent Breakthroughs in Natural Language Processing

Stanford

Christopher Manning

Director, Stanford Artificial Intelligence Laboratory

@chrmanning ❀ @stanfordnlp

BAAI 2019

“the common misconception [is] that language use has primarily to do with words and what they mean.

It doesn’t. It has primarily to do with people and what *they* mean.”

Asking questions and influencing answers

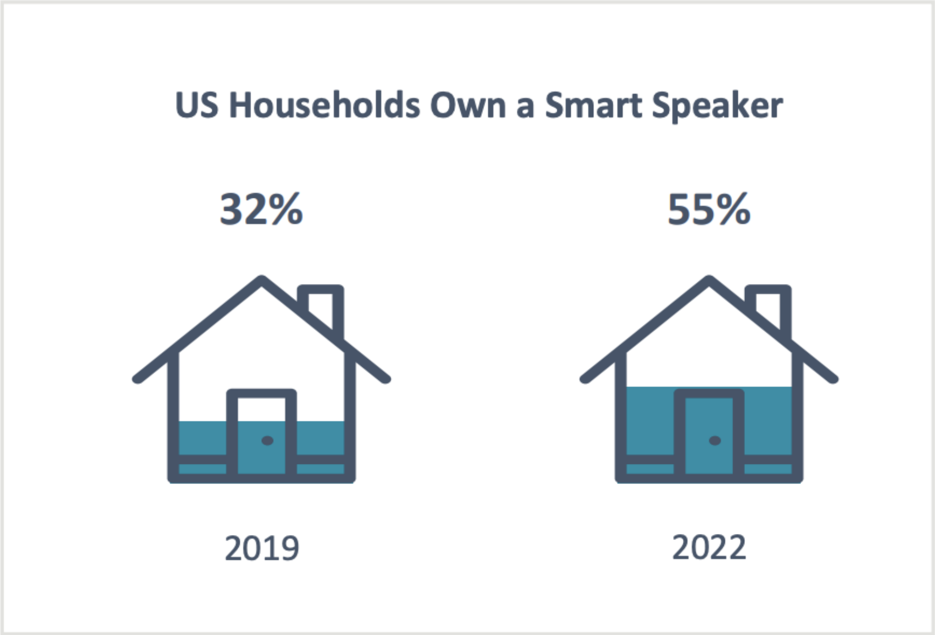
Clark & Schober, 1992

9:41
Tuesday, September 12

By 2020, 40% of users will be interacting with primarily new applications that support conversational UIs with Artificial Intelligence

Source: Gartner’s Competitive Landscape 2018

“Smart speaker” virtual assistants



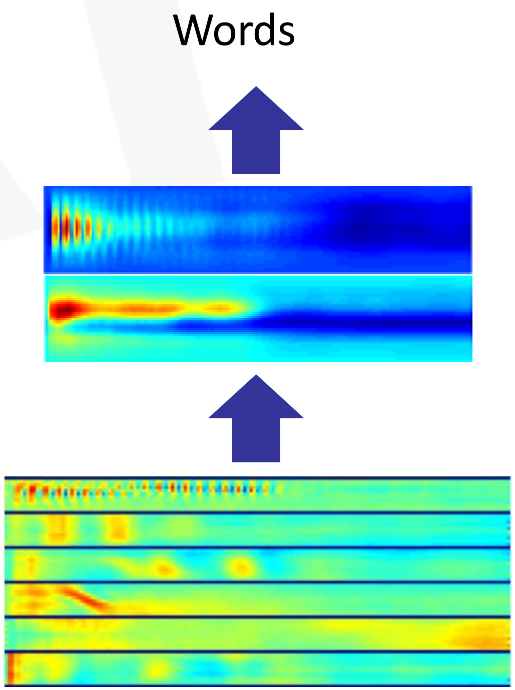
Speech recognition is now 3 times as fast as texting!

(Ruan, Landay, and Ng 2016)

YouTube

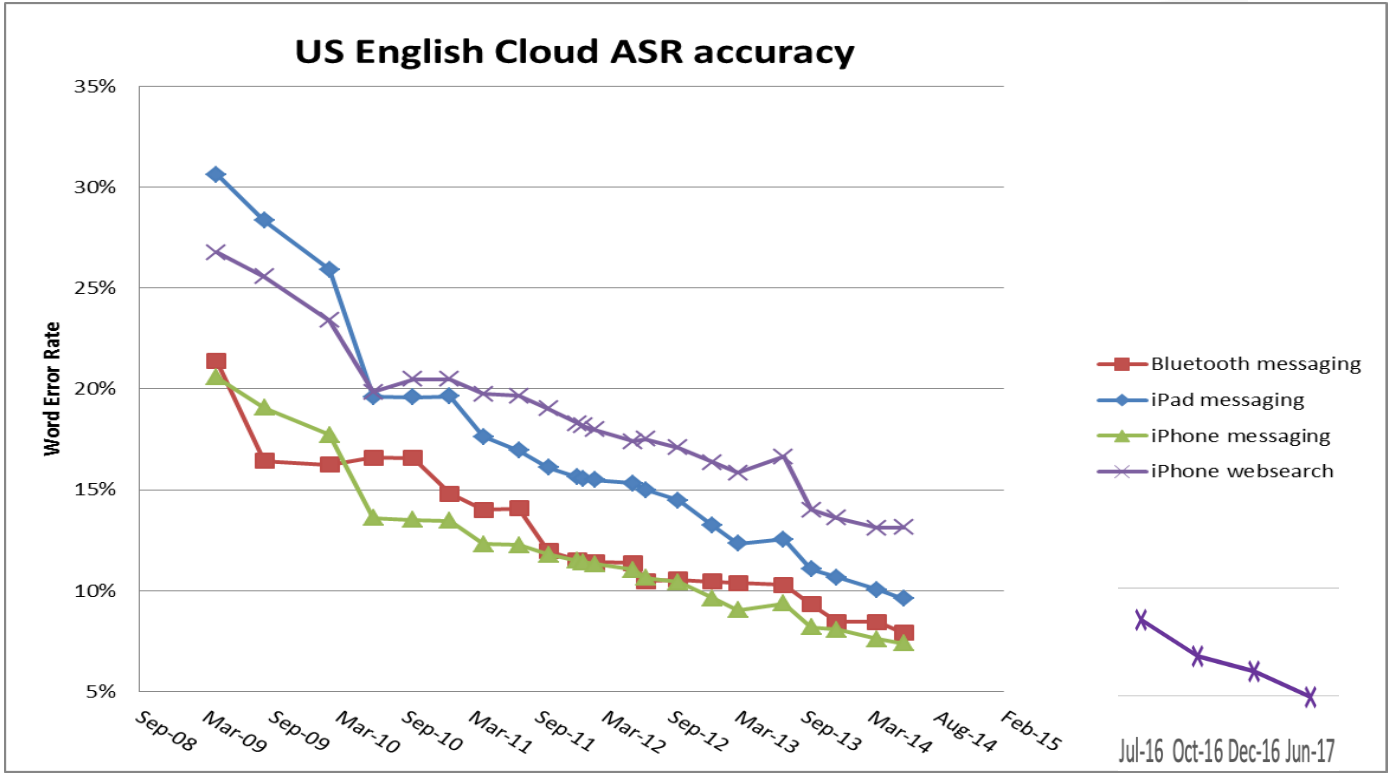
Deep Learning for Speech Recognition

- The first breakthrough results of “deep learning” on large datasets happened in **speech recognition**
- George Dahl et al. (2010/2012): Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition



Acoustic model \ WER	RT03S FSH	Hub5 SWB
Traditional GMM (D. et al. 2012)	27.4	23.6
Deep Learning (Dahl et al. 2012)	18.5 (−33%)	16.1 (−32%)
Deep Learning (Saon et al. 2017)	8.0 (−71%)	5.5 (−77%)

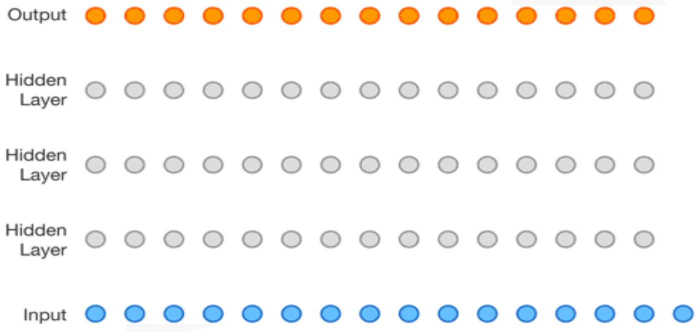
ASR accuracy improvements



Deep Learning for Generation of Speech

WaveNet: A deep generative model of raw audio

DeepMind (van den Oord et al. 2016) https://arxiv.org/abs/1609.03499



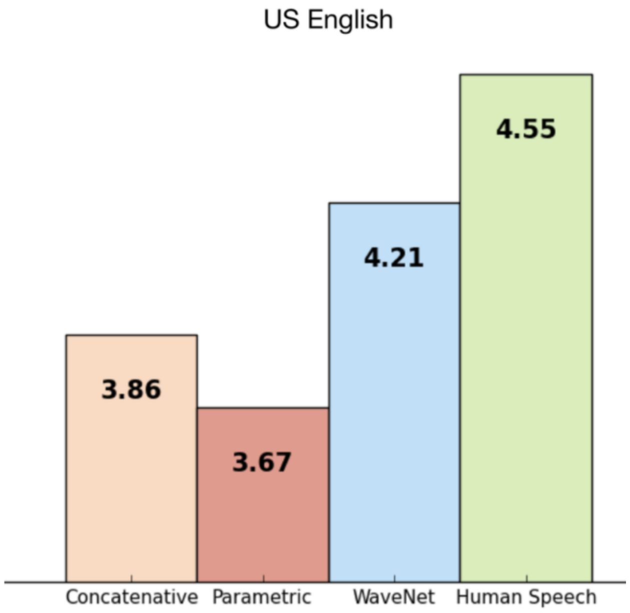
Old (CMU c.2000)

Concatenative

Parametric

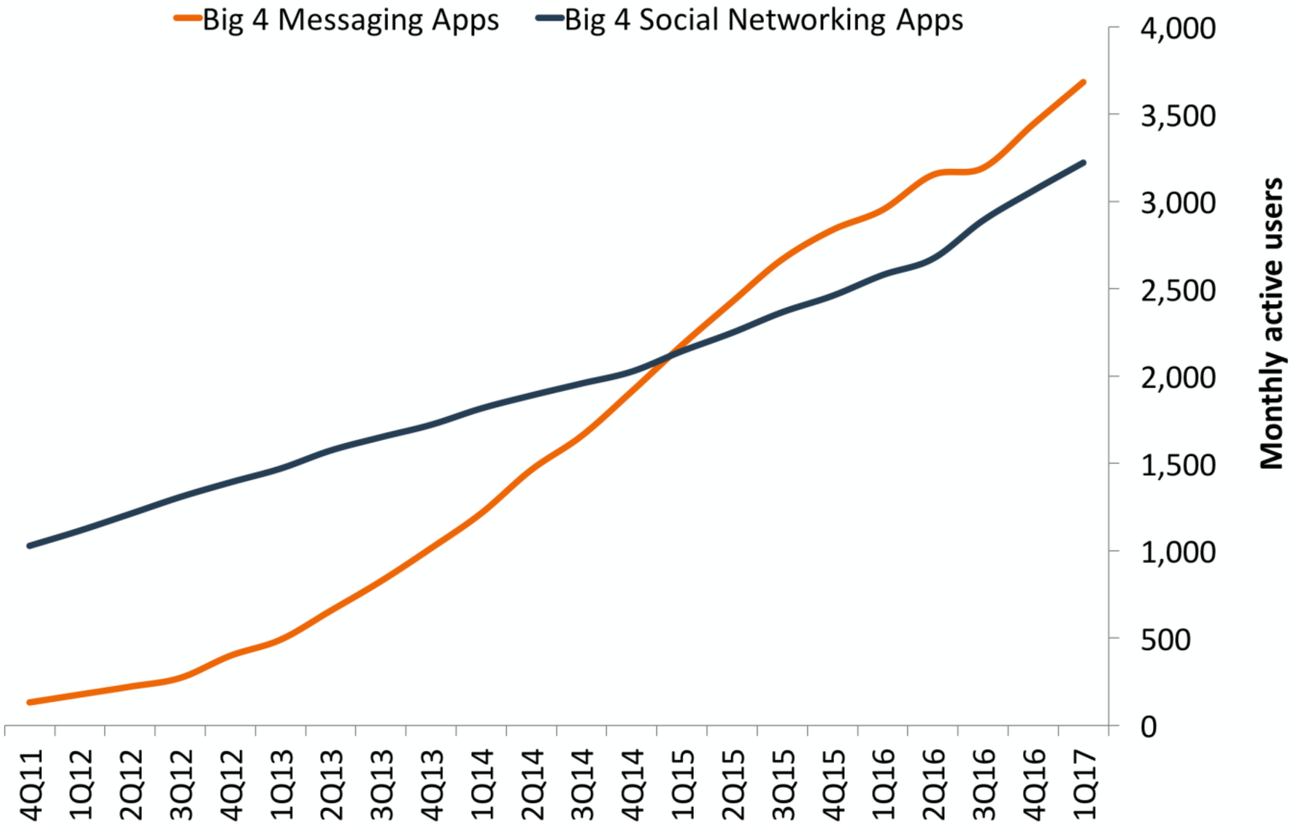
WaveNet

Quality: Mean Opinion Scores



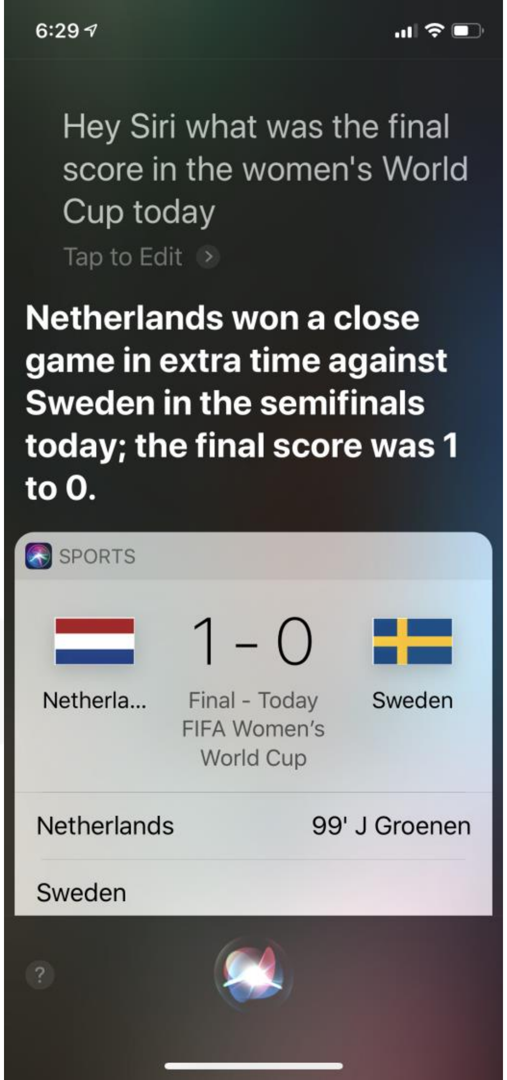
Messaging Apps Have Surpassed Social Networks

Global monthly active users for the top 4 messaging apps and social networks, In millions



Note: Big 4 messaging apps are WhatsApp, Messenger, WeChat, Viber.
Big 4 social networks are Facebook, Instagram, Twitter, LinkedIn
Source: Companies, Apptopia, TechCrunch, BI Intelligence estimates, 2017

BI INTELLIGENCE



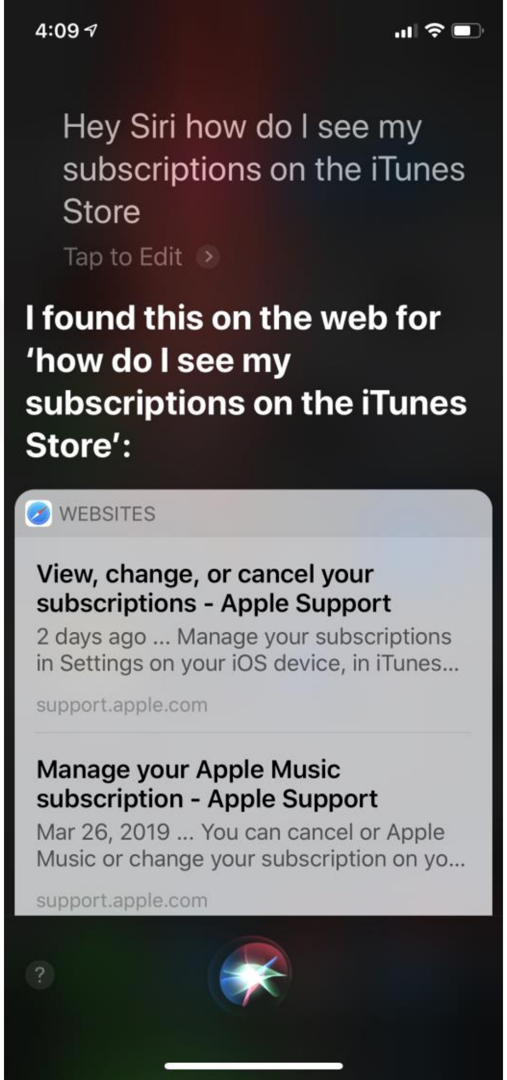
Hey Siri what was the final score in the women's World Cup today

Netherlands won a close game in extra time against Sweden in the semifinals today; the final score was 1 to 0.

SPORTS

Netherla...	1 - 0 Final - Today FIFA Women's World Cup	Sweden

Netherlands	99' J Groenen
Sweden	



Hey Siri how do I see my subscriptions on the iTunes Store

I found this on the web for 'how do I see my subscriptions on the iTunes Store':

WEBSITES

View, change, or cancel your subscriptions - Apple Support
2 days ago ... Manage your subscriptions in Settings on your iOS device, in iTunes...
support.apple.com

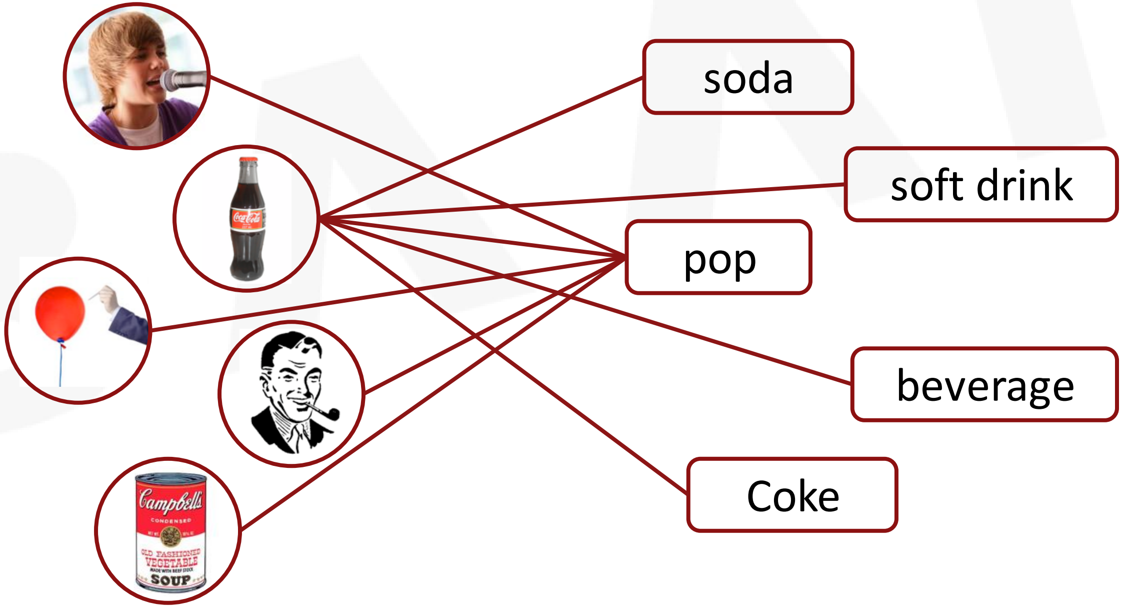
Manage your Apple Music subscription - Apple Support
Mar 26, 2019 ... You can cancel or Apple Music or change your subscription on yo...
support.apple.com

Why is human language hard to understand?

Human languages:

- Are highly ambiguous at all levels
- Are fuzzy and vague
- Require reasoning about the world to understand
- Exploit context to convey meaning
- Use features like recursive structures and coreference
- Are part of a social system of persuading, insulting, amusing, …

Meaning and reference



soda

soft drink

pop

beverage

Coke

OK, why *else* is NLP hard? Many reasons!

non-standard language

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
retweet
bromance
teabagger

garden path sentences

The man who hunts ducks out on weekends.

The cotton shirts are made from grows here.

tricky entity names

… a mutation on the for gene …

Where is A Bug's Life playing …

Most of Let It Be was recorded …

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

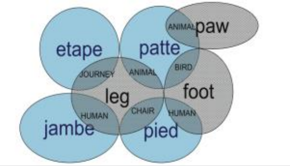
prosody

I never said ***she*** stole my money.

I never said she ***stole*** my money.

I never said she stole ***my*** money.

lexical specificity



“(Artificial) neural (network)” or “deep learning” models for word meaning

We represented a word as a vector of numbers

$$\text{versatile} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Similar vectors = similar meaning

Learn vectors via distributional similarity

"You shall know a word by the company it keeps" (J. R. Firth 1957: 11)

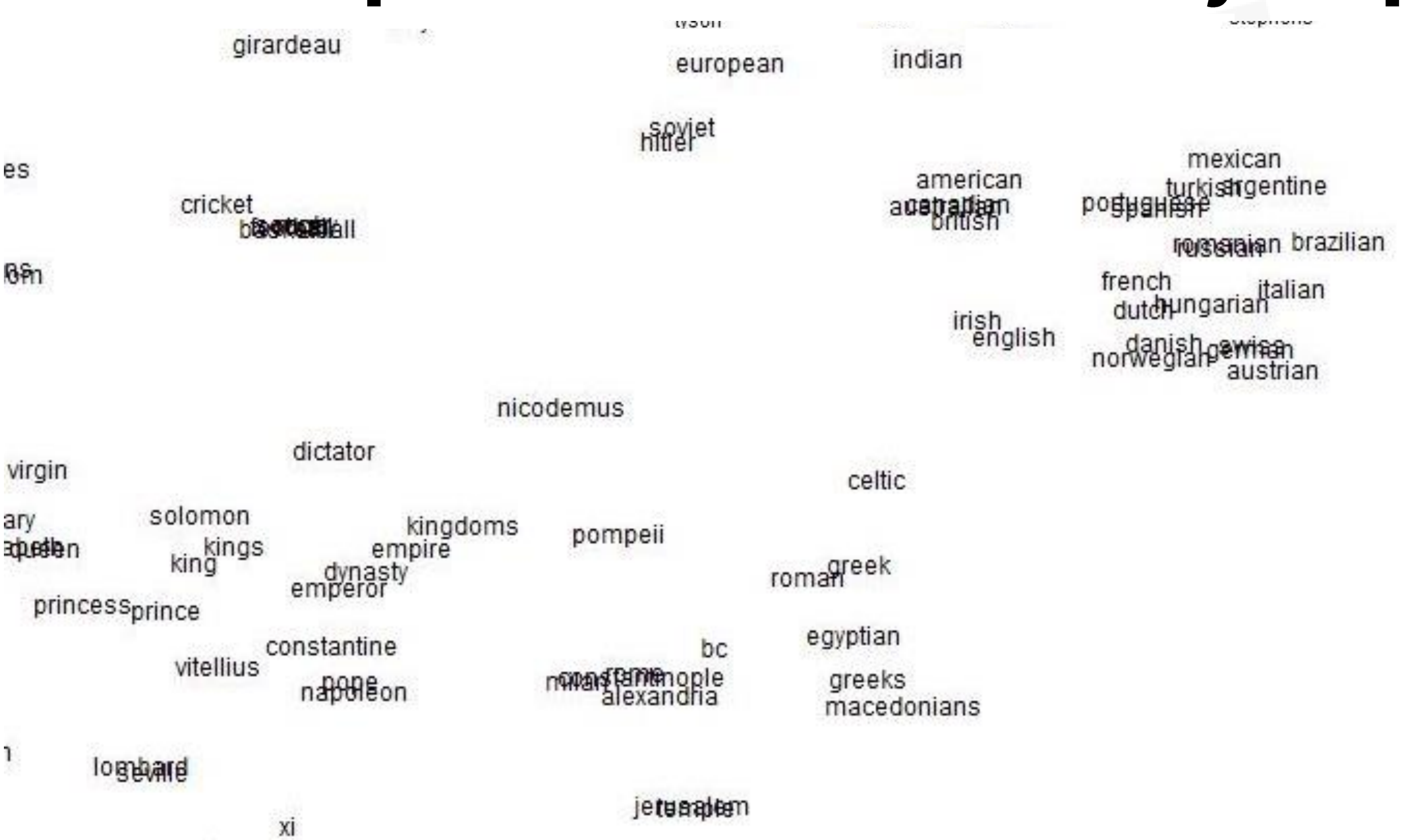
Defining similarity via contextual distributions in texts is one of the most successful ideas of modern computational linguistics

any devices with a web browser, from laptops and tablets to smart phones

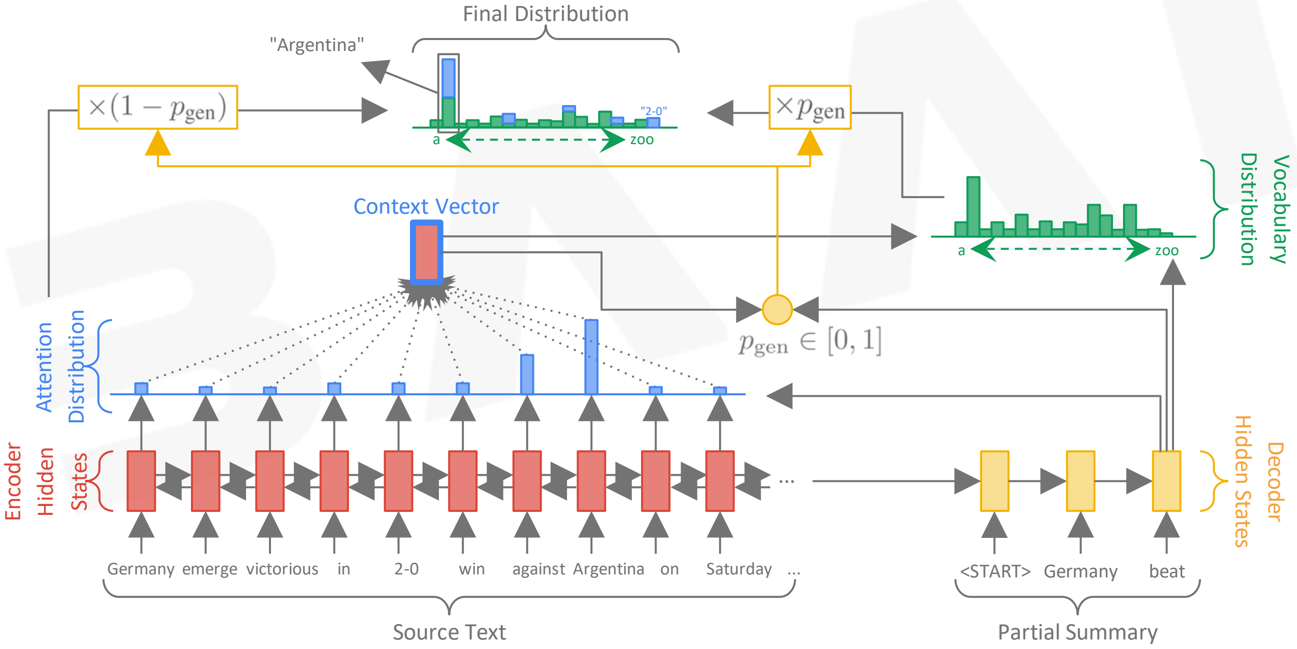
Users can download it for home computers or laptops from Microsoft Update website

↖ These words will represent ***laptops*** ↗

Word vectors put similar words nearby in space



Vectors go into (complex!) neural net systems



Neural pointer-generator network for text summarization
See, Liu, and Manning 2018

Natural language is key to many applications

> “The role of FAIR (Facebook AI Research) is to advance the science and the technology of AI and do experiments that demonstrate that technology for new applications like computer vision, dialogue systems, virtual assistants, speech recognition, natural language understanding, translation, things like that.”

Yann LeCun

http://www.businessinsider.com/interview-yann-lecun-facebook-ai-deepmind-2016-10

Commercial uses of NLP

Tasks

- Market Intelligence / Competitive advantage
- Understanding customer experience
- Brand perception
- Hiring (advertising; filtering resumes)
- Customer service/support
- Sales / Managing the sales funnel
- Finance: Trading on information
- E-commerce

Technologies

- Information (fact) extraction
- Sentiment analysis
- Semantic Search
- Question Answering
- Chatbots
- Neural Machine Translation
- Opinion mining

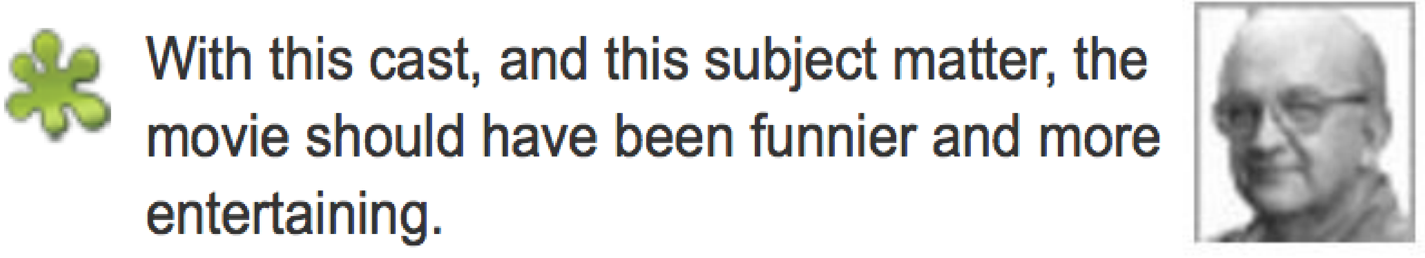
Sentiment analysis

Is someone expressing positive, negative, or neutral views?

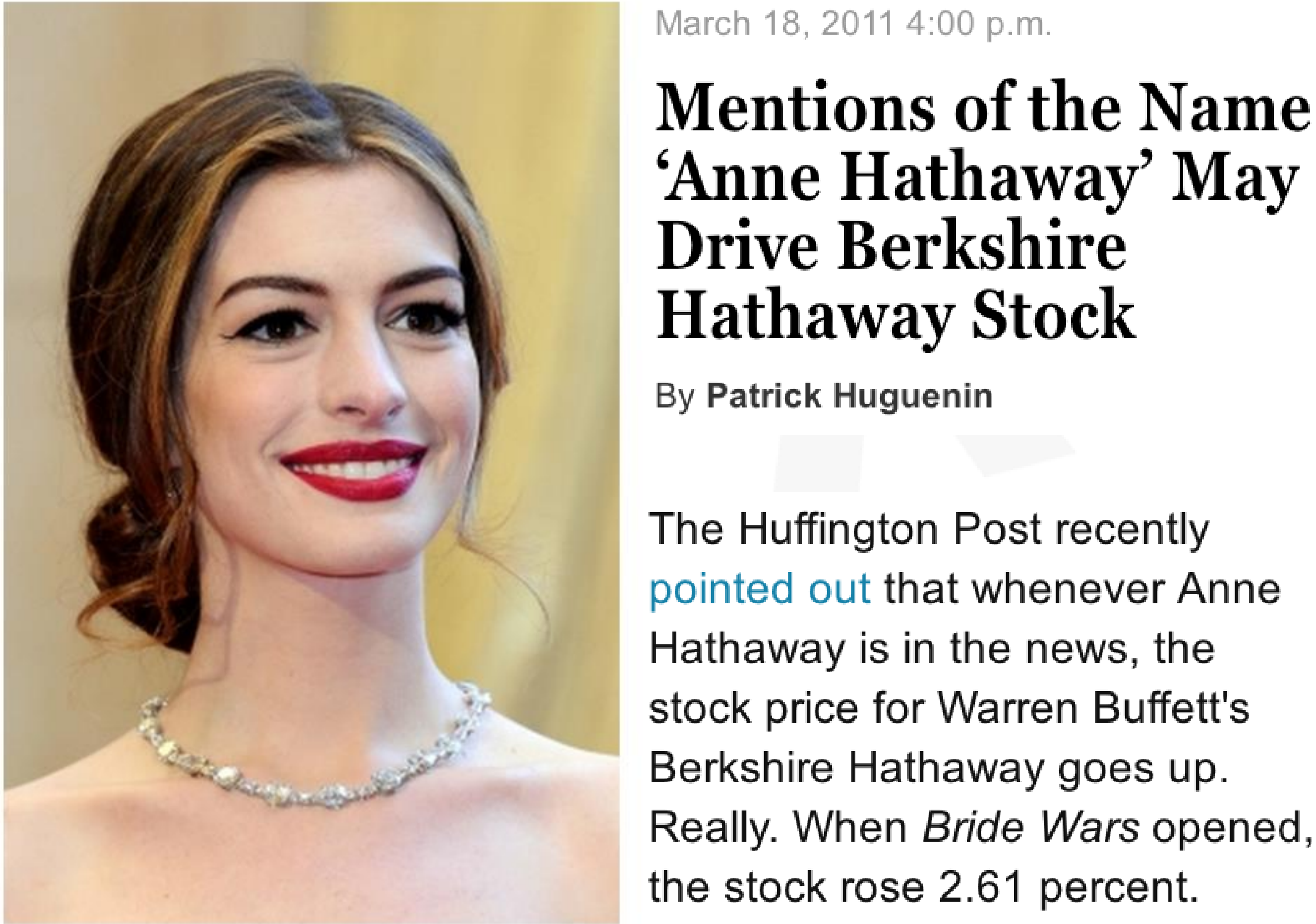
Sometimes easy looking at a “bag of words”

… loved … … … … … great … … … … … impressed … … entertaining …

But often it’s more subtle



With this cast, and this subject matter, the movie should have been funnier and more entertaining.



March 18, 2011 4:00 p.m.

Mentions of the Name 'Anne Hathaway' May Drive Berkshire Hathaway Stock

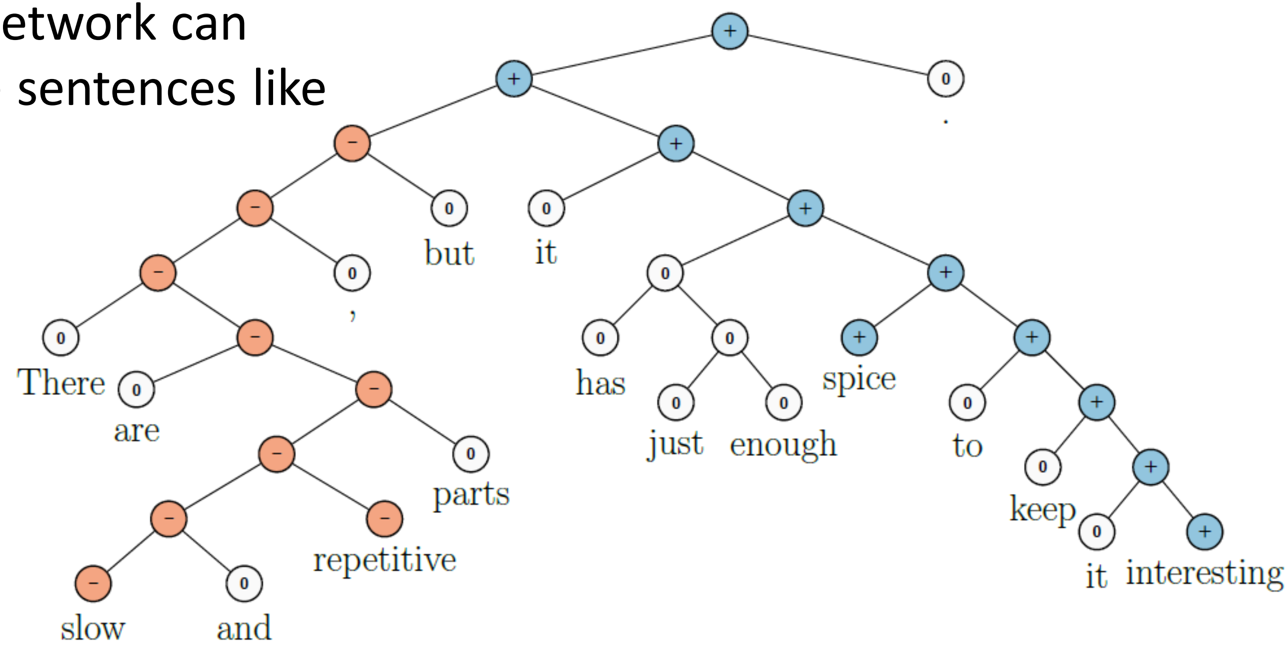
By **Patrick Huguenin**

The Huffington Post recently pointed out that whenever Anne Hathaway is in the news, the stock price for Warren Buffett's Berkshire Hathaway goes up. Really. When *Bride Wars* opened, the stock rose 2.61 percent.

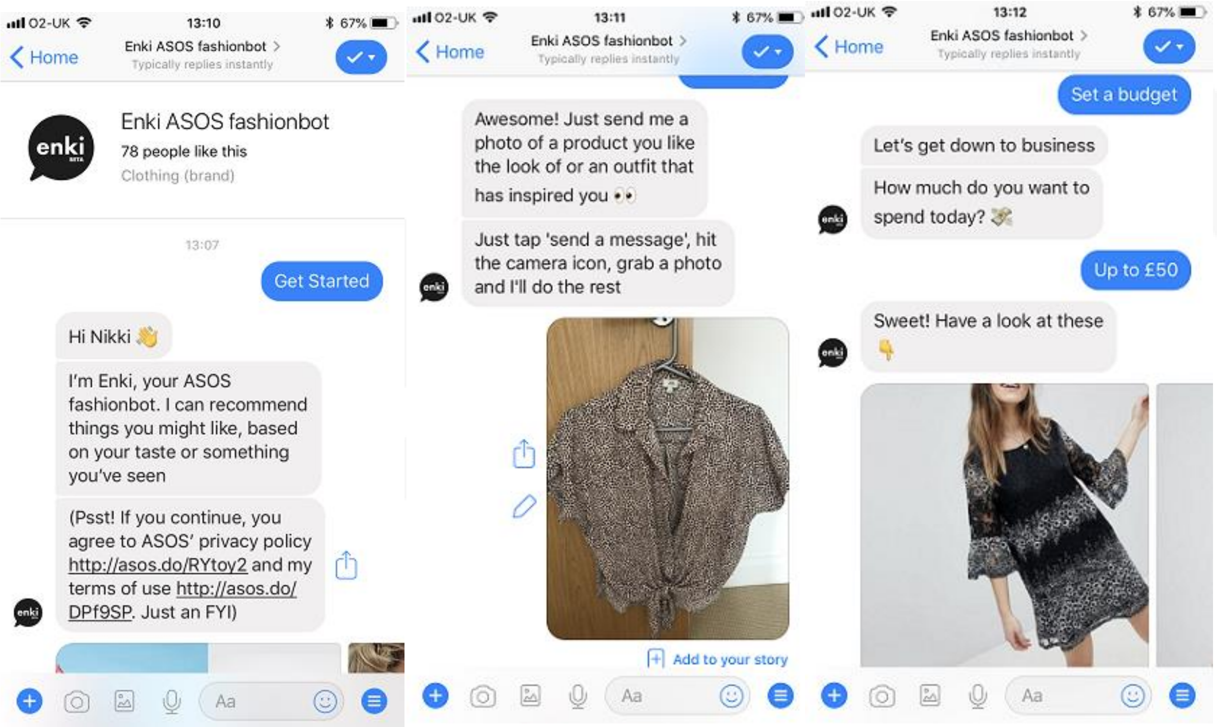
Tree-structured Neural Sentiment Analysis

[Tai, Socher & Manning 2015]

A tree-structured network can capture contrastive sentences like X but Y



Fashion retailer ASOS's chatbot Enki



Hi Nikki 👋

I'm Enki, your ASOS fashionbot. I can recommend things you might like, based on your taste or something you've seen

(Psst! If you continue, you agree to ASOS' privacy policy http://asos.do/RYtoy2 and my terms of use http://asos.do/DPf9SP. Just an FYI)

Awesome! Just send me a photo of a product you like the look of or an outfit that has inspired you 👀

Just tap 'send a message', hit the camera icon, grab a photo and I'll do the rest

Let's get down to business

How much do you want to spend today? 💸

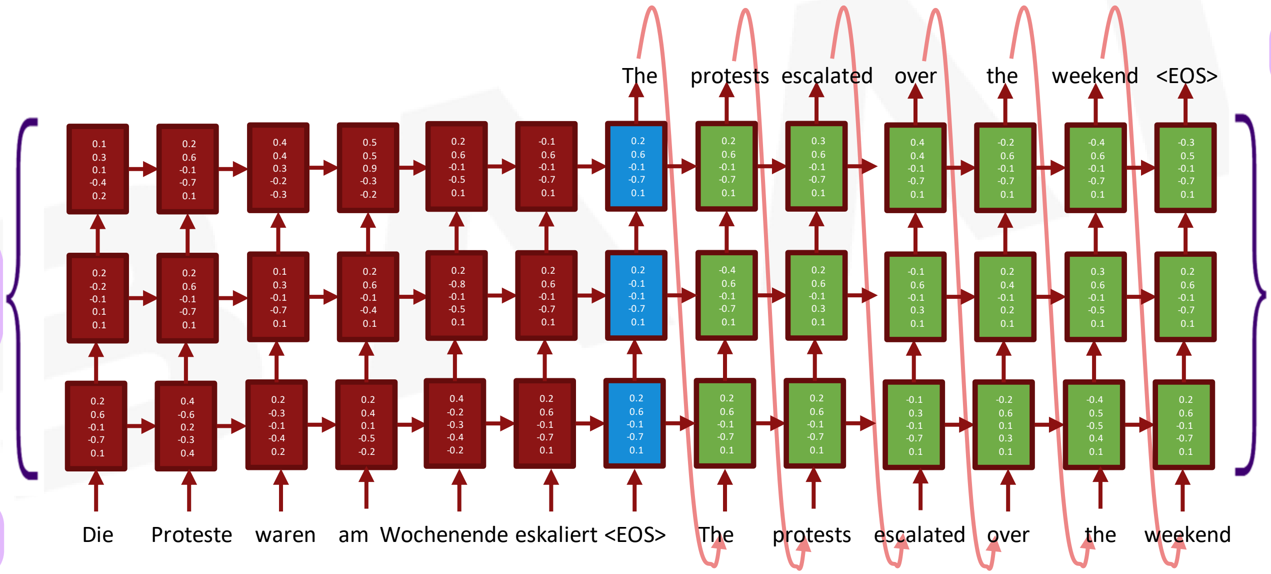
Sweet! Have a look at these 👇

- 35% more people reached
- 300% increase in orders
- 250% increase in return on ad spend
  - Vs. previous pretty underwhelming "gift assistant"

https://www.singlegrain.com/artificial-intelligence/effects-of-natural-language-processing-nlp-on-digital-marketing/

Fashion retailer ASOS’s chatbot Enki

Neural network for machine translation



Encoder: Builds up sentence meaning

Source sentence

Translation generated

Decoder

Feeding in last word

Enormously fast and successful transition of Neural Machine Translation to commercial use!

2014: First modern research attempts on neural MT

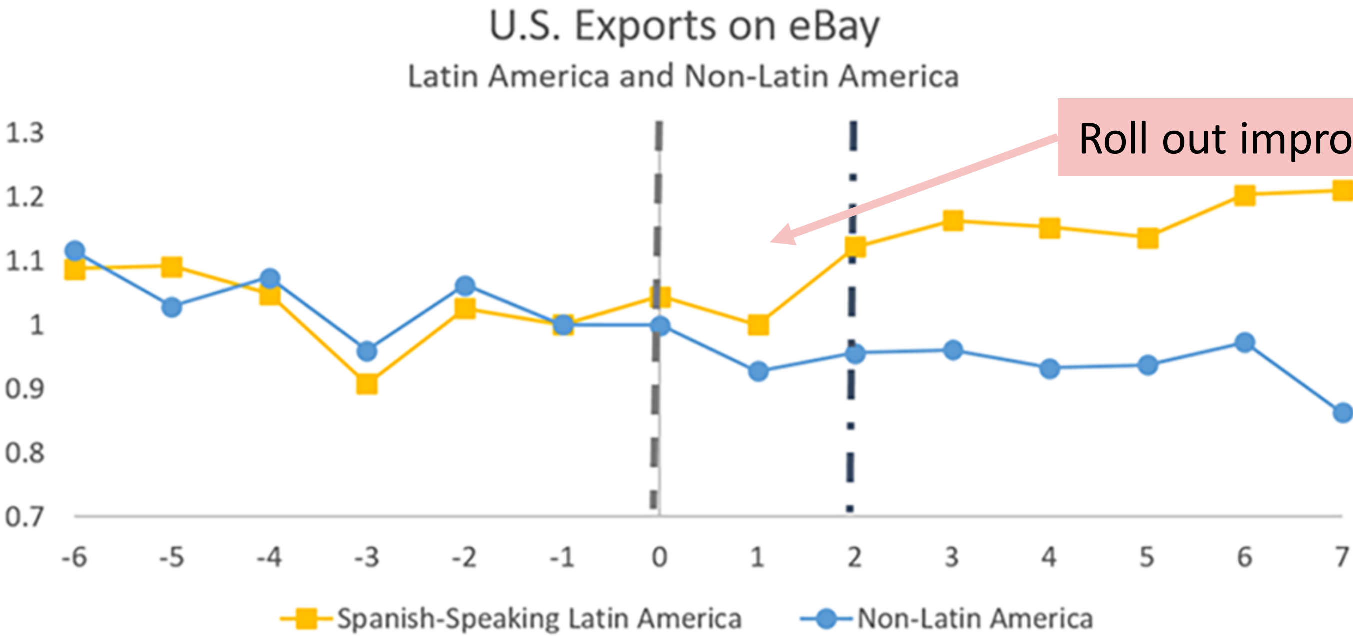
At Google, U. Montréal, and Stanford

2017+: Almost everyone is using Neural MT in production, for many/all language pairs

Microsoft SYSTRAN beyond language Google facebook

Baidu 百度 網易 NetEase www.163.com Tencent 腾讯 搜狗搜索

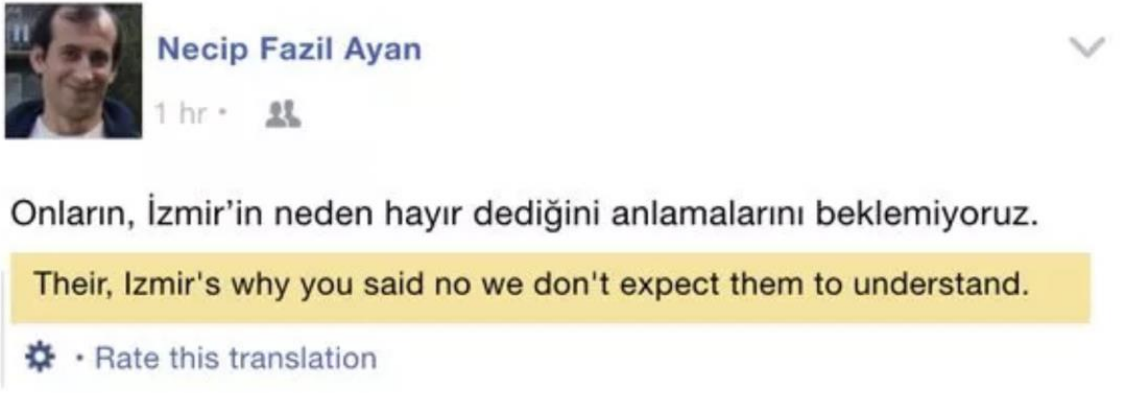
Better translation boosted eBay sales more than 10%



Roll out improved MT

Erik Brynjolfsson et al. MIT 2018.
https://www.theverge.com/2019/5/15/18624459/ai-translation-boosted-ebay-sales-more-than-10-percent

Phrase-based Statistical MT



Necip Fazil Ayan

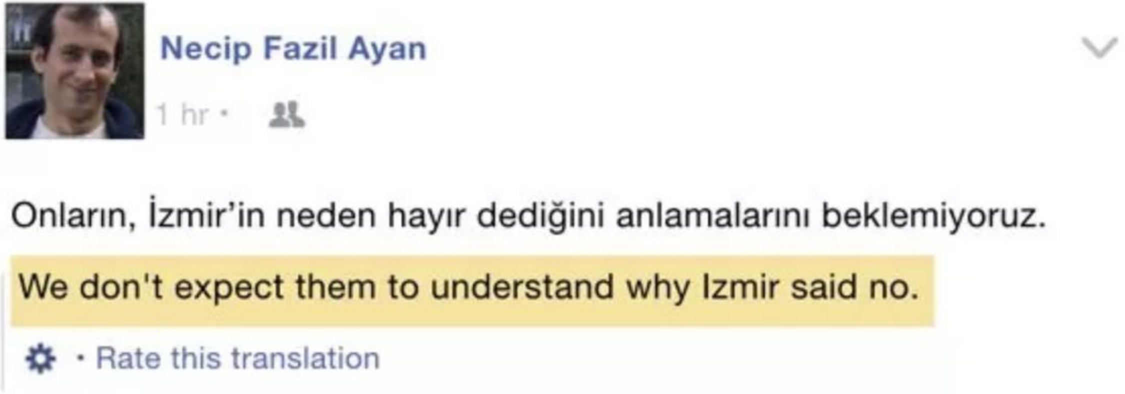
1 hr ·

Onların, İzmir'in neden hayır dediğini anlamalarını beklemiyoruz.

Their, Izmir's why you said no we don't expect them to understand.

· Rate this translation

Neural Machine Translation



Necip Fazil Ayan

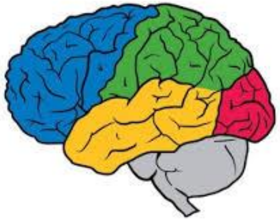
1 hr ·

Onların, İzmir'in neden hayır dediğini anlamalarını beklemiyoruz.

We don't expect them to understand why Izmir said no.

· Rate this translation

ELECTRA: Efficiently Pre-training Text Encoders as Discriminators rather than Generators



Kevin Clark

Thang Luong, Quoc Le, Chris Manning

NLP: 2018 breakthrough with big language models

ULMfit
Jan 2018
Training:
103M words
1 GPU day

fast.ai

All of these models are Transformer models

GPT
June 2018
Training
800M words
240 GPU days

OpenAI

BERT
Oct 2018
Training
3.3B words
256 TPU days
~320–560
GPU days

Google AI

GPT-2
Feb 2019
Training
40B words
~2048 TPU v3 days
according to [a reddit thread]

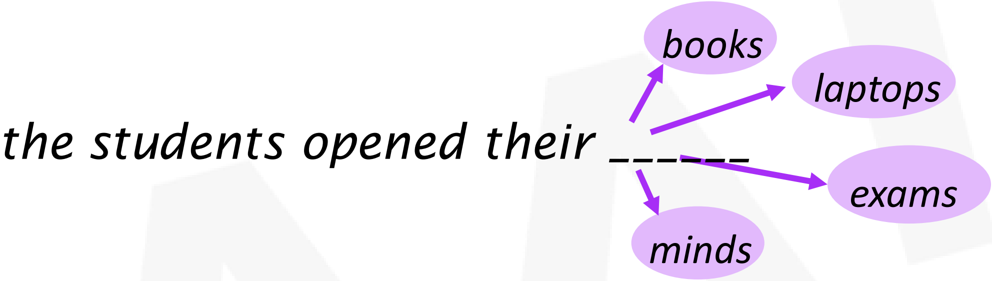
OpenAI

XL-Net,
ERNIE,
Grover
RoBERTa, …
July 2019

Facebook AI Research (FAIR)
Google AI
UWNLP
Baidu 百度
Carnegie Mellon University

Language Modeling

A **Language Model (LM)** predicts a word in a context



For a word sequence $x^{(1)}, x^{(2)}, \dots, x^{(t)}$, it gives the probability of the next word $x^{(t+1)}$:

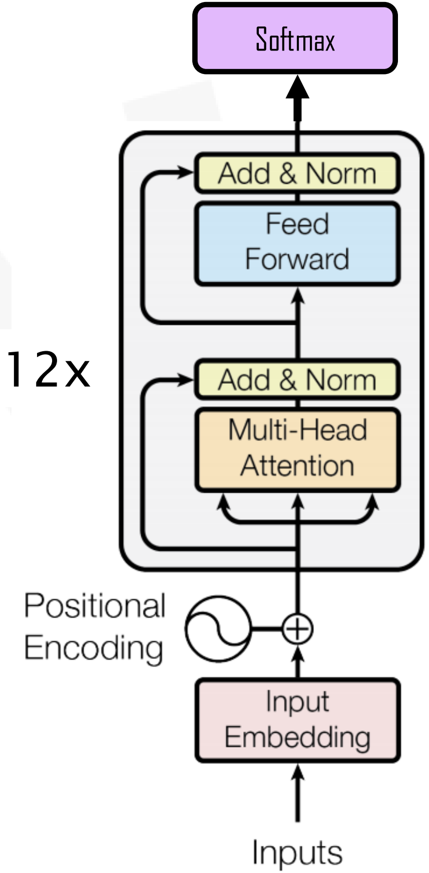
$$P(x^{(t+1)} | x^{(t)}, \dots, x^{(1)})$$

An LM is a key part of decoding tasks like **speech recognition, spelling correction**, and any NL generation task, including **machine translation, summarization**, and **story generation**

Transformer (Vaswani et al. 2017)

https://arxiv.org/pdf/1706.03762.pdf

- **Non-recurrent** sequence model (or sequence-to-sequence model)
- A **deep** model with a sequence of **attention-**based transformer blocks
- Depth allows a certain amount of lateral information transfer in understanding sentences, in slightly unclear ways
- Final cost/error function is standard cross-entropy error on top of a softmax classifier



GPT-2 language model (cherry-picked) output

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

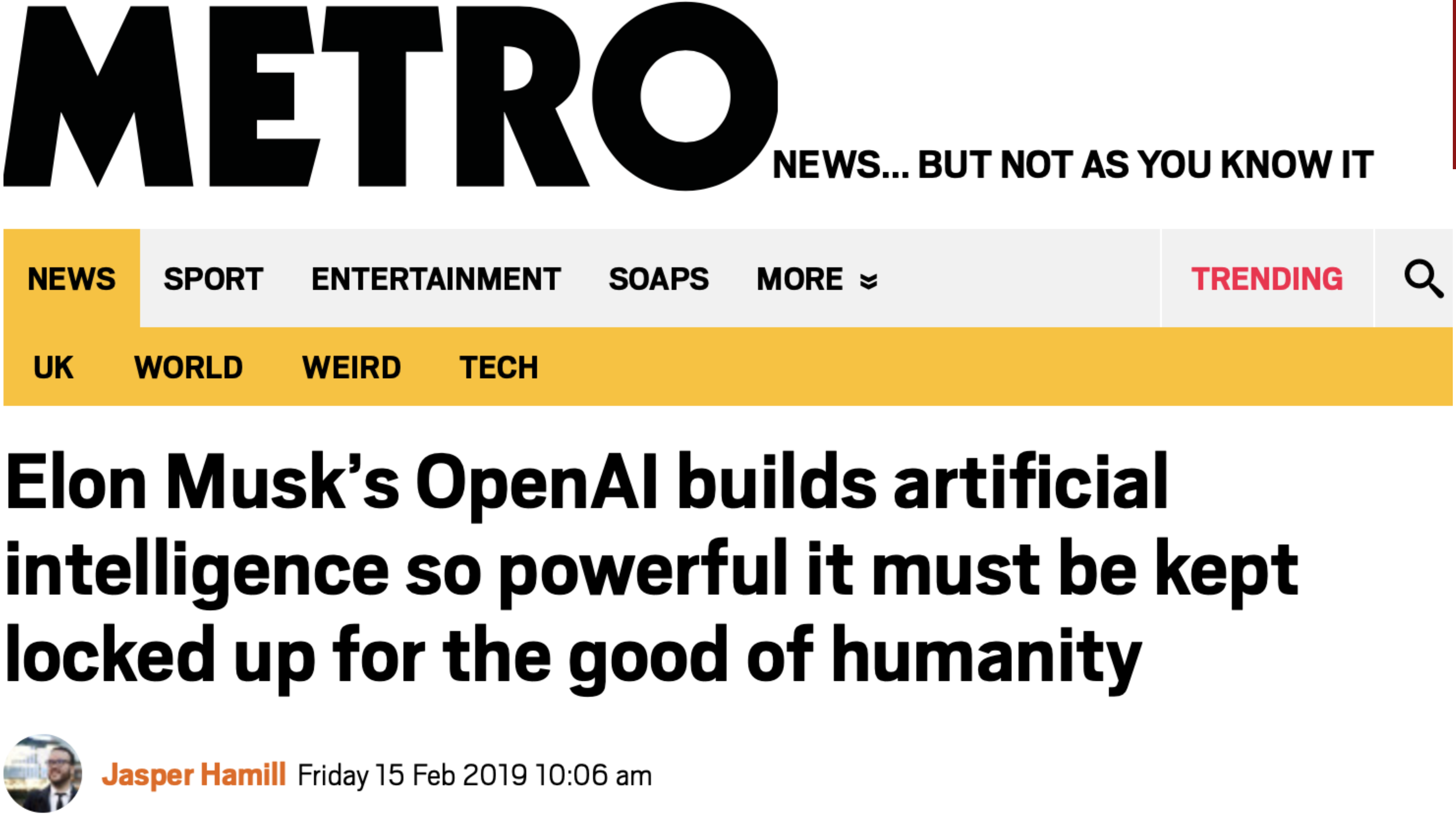
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

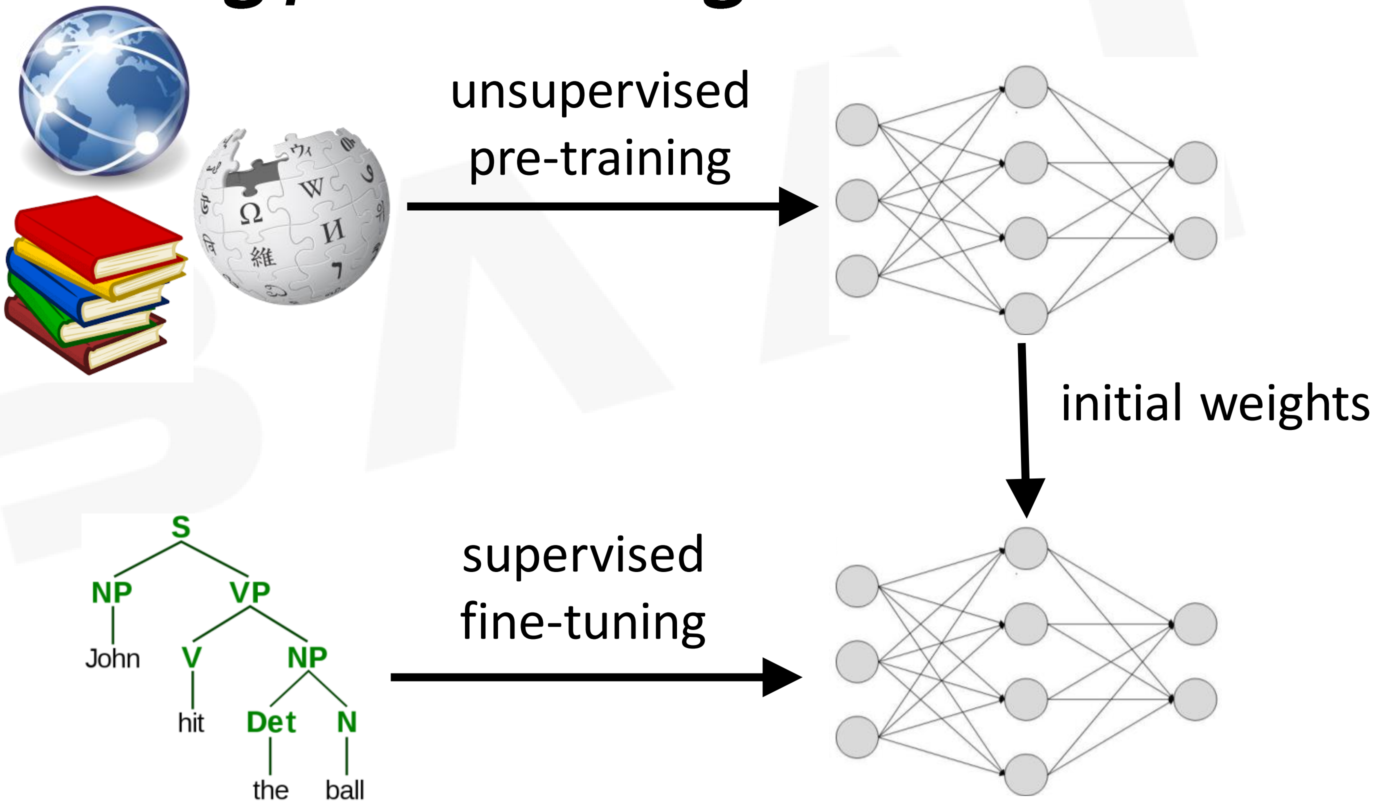


METRO NEWS... BUT NOT AS YOU KNOW IT

Elon Musk’s OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

Jasper Hamill Friday 15 Feb 2019 10:06 am

Pre-training / Fine-tuning



unsupervised pre-training

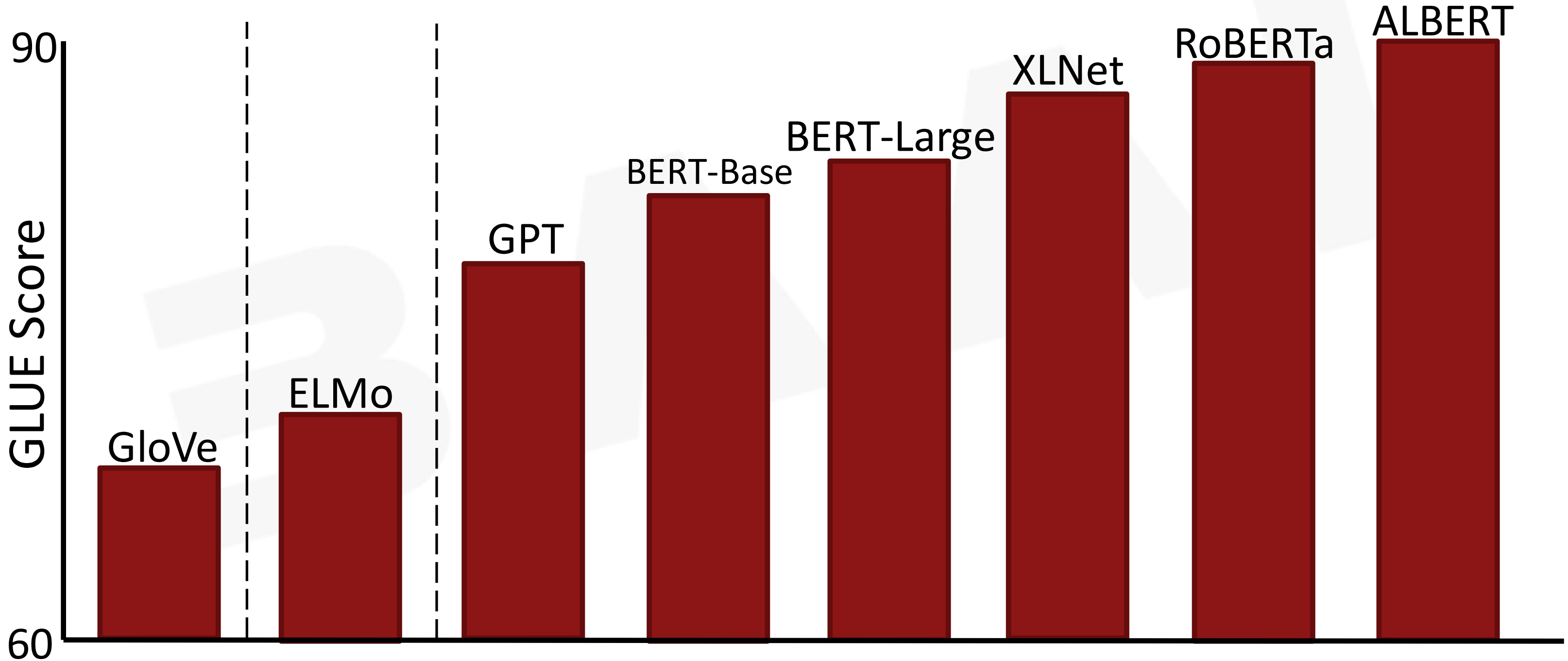
initial weights

supervised fine-tuning

GLUE tasks for Natural Language Understanding

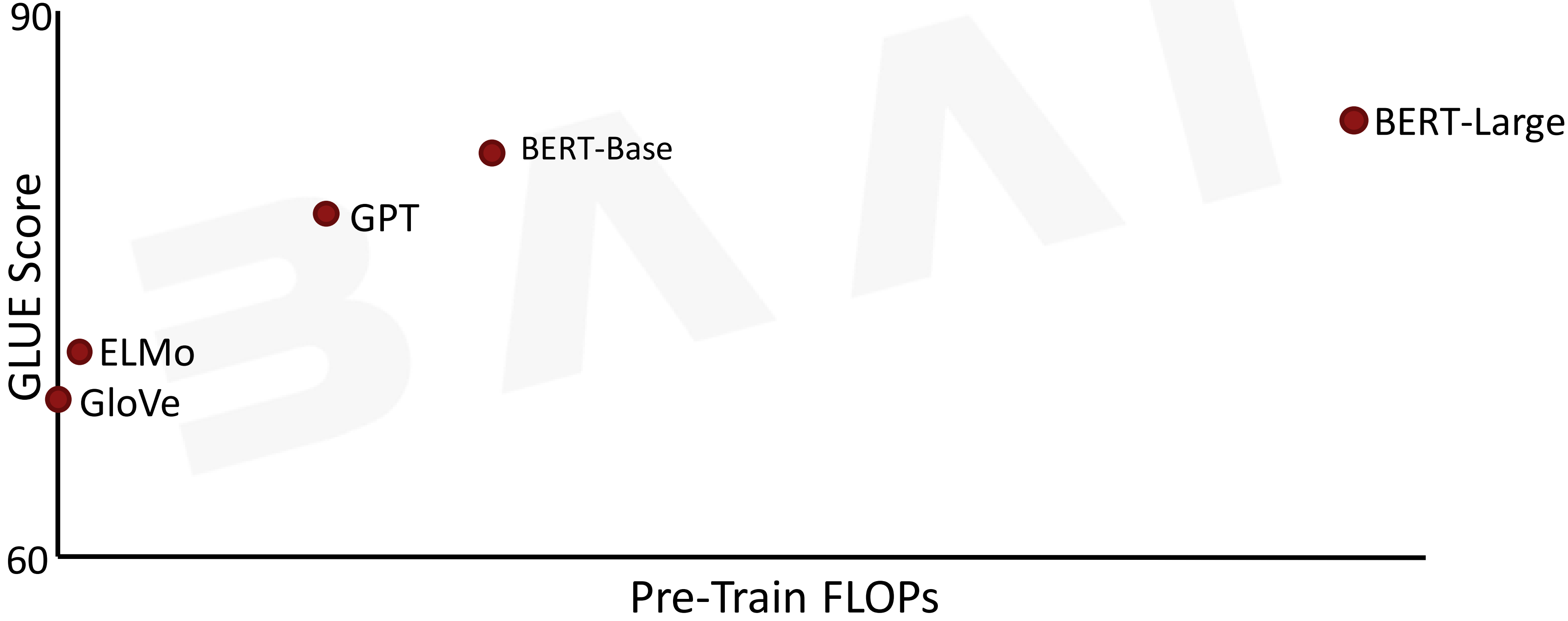
- GLUE benchmark (Wang et al. ICLR 2018) is dominated by natural language inference tasks, but also has sentence similarity, sentiment, linguistic acceptability, Winograd schema
- **MultiNLI**
- Premise: Hills and mountains are especially sanctified in Jainism.
  Hypothesis: Jainism hates nature. Label: Contradiction
- **CoLa**
- Sentence: The wagon rumbled down the road. Label: Acceptable
- Sentence: The car honked down the road. Label: Unacceptable

Rapid Progress from Pre-Training



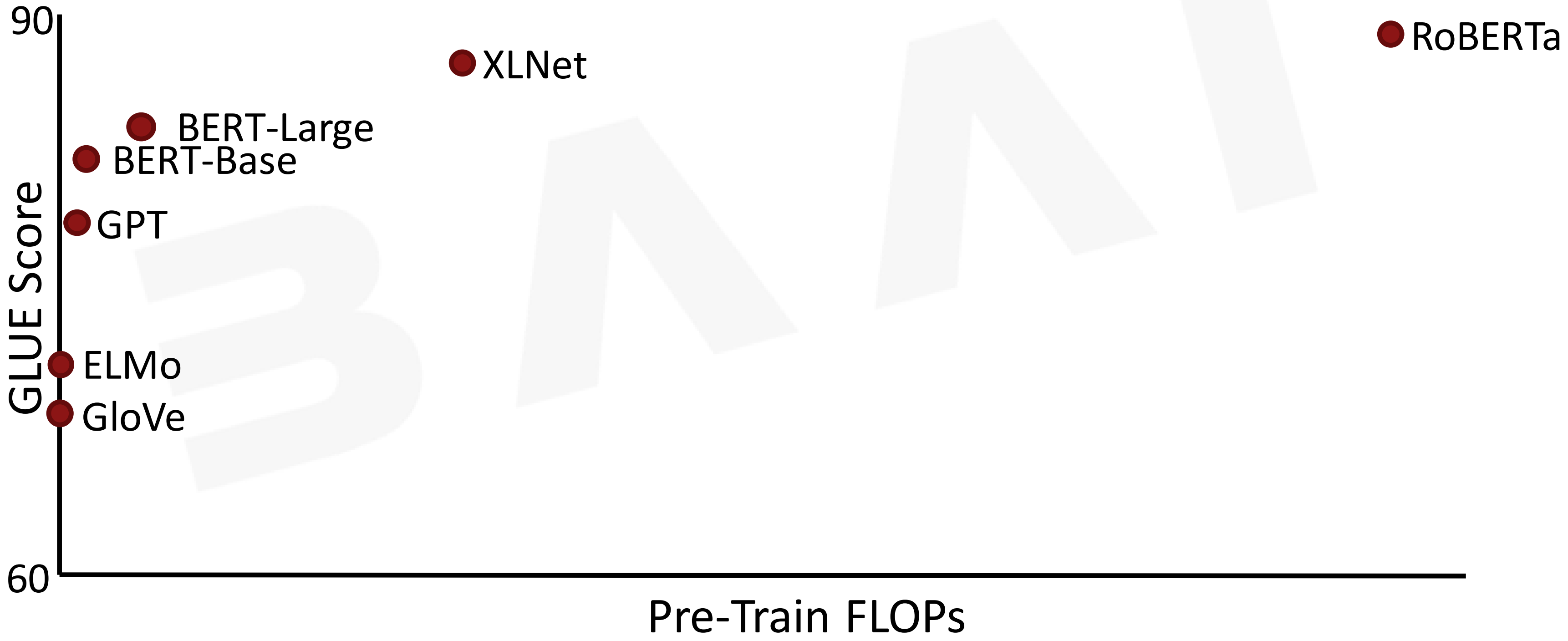
Over 3x reduction in error since 2017, “superhuman” performance

But let’s change the x-axis to compute …



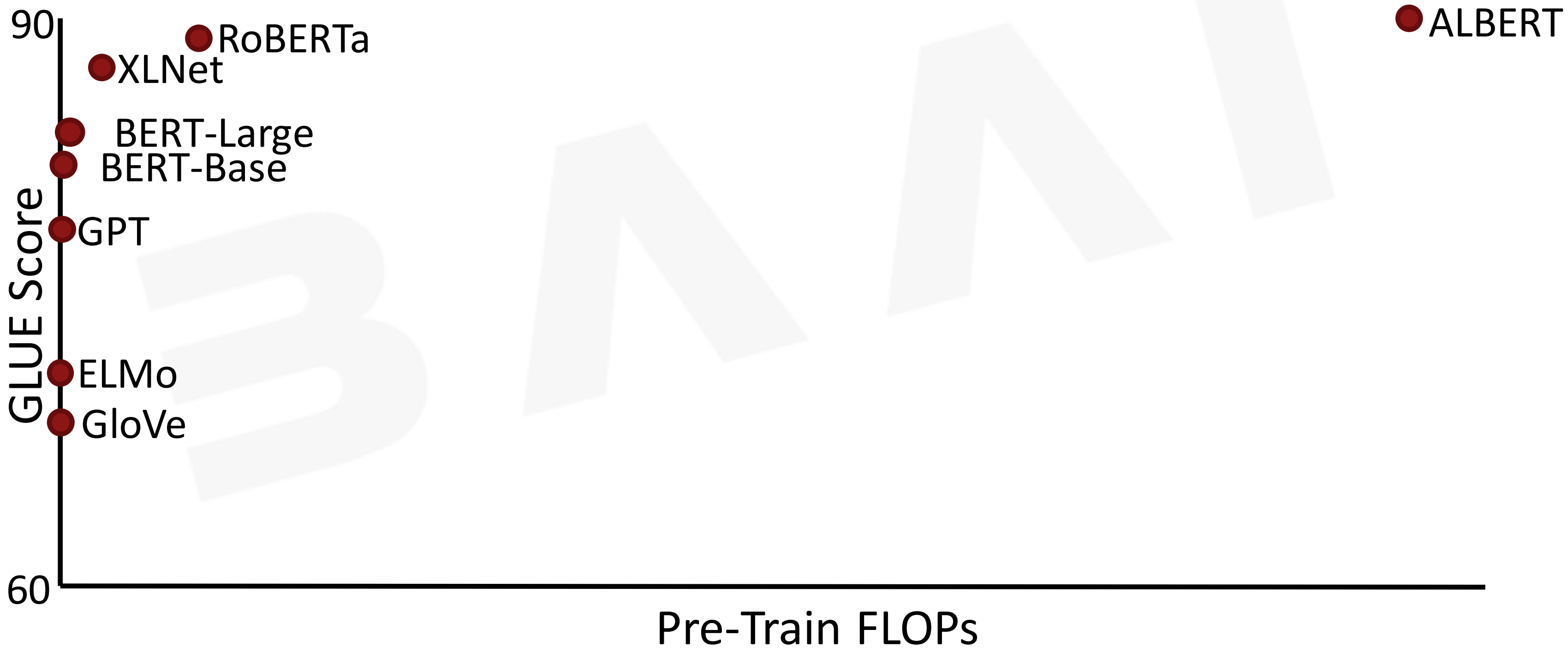
BERT-Large uses 60x more compute than ELMo

But let’s change the x-axis to compute …



RoBERTa uses 16x more compute than BERT-Large

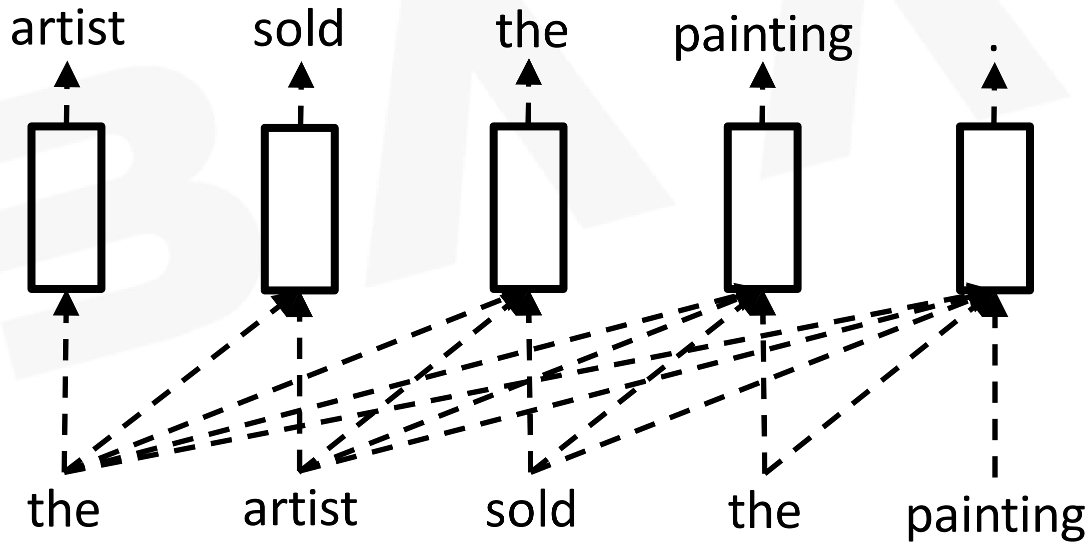
More compute more better?



ALBERT uses 10x more compute than RoBERTa

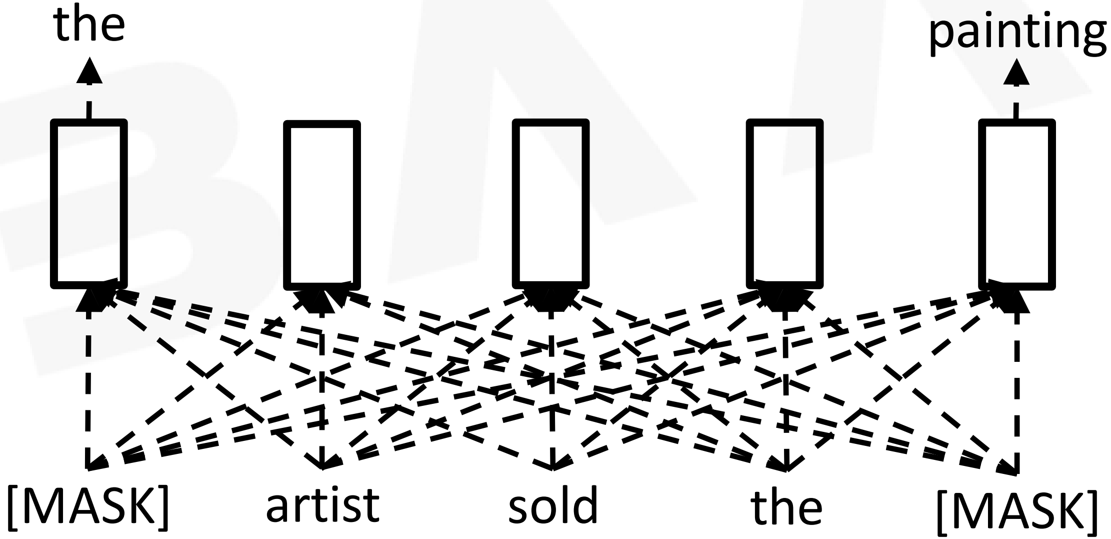
Language Model Pretraining

- ULMFit, ELMo, GPT, …



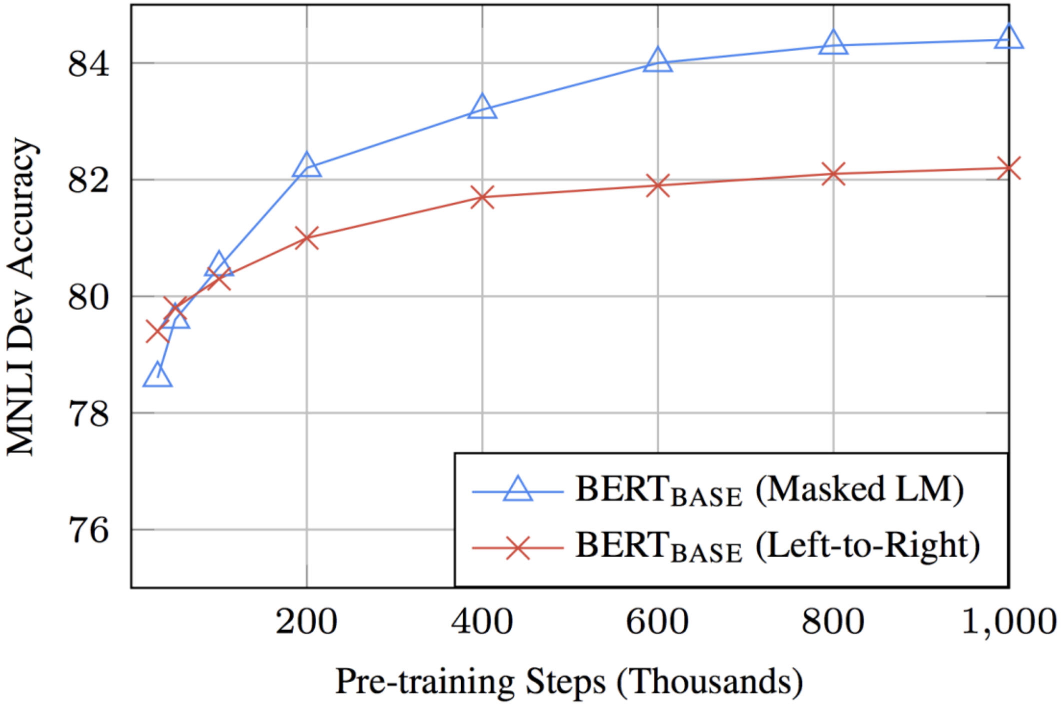
Masked Language Model Pretraining

- BERT, XLNet, RoBERTa, …



Masked Language Model Pretraining

- Bidirectional gives better performance

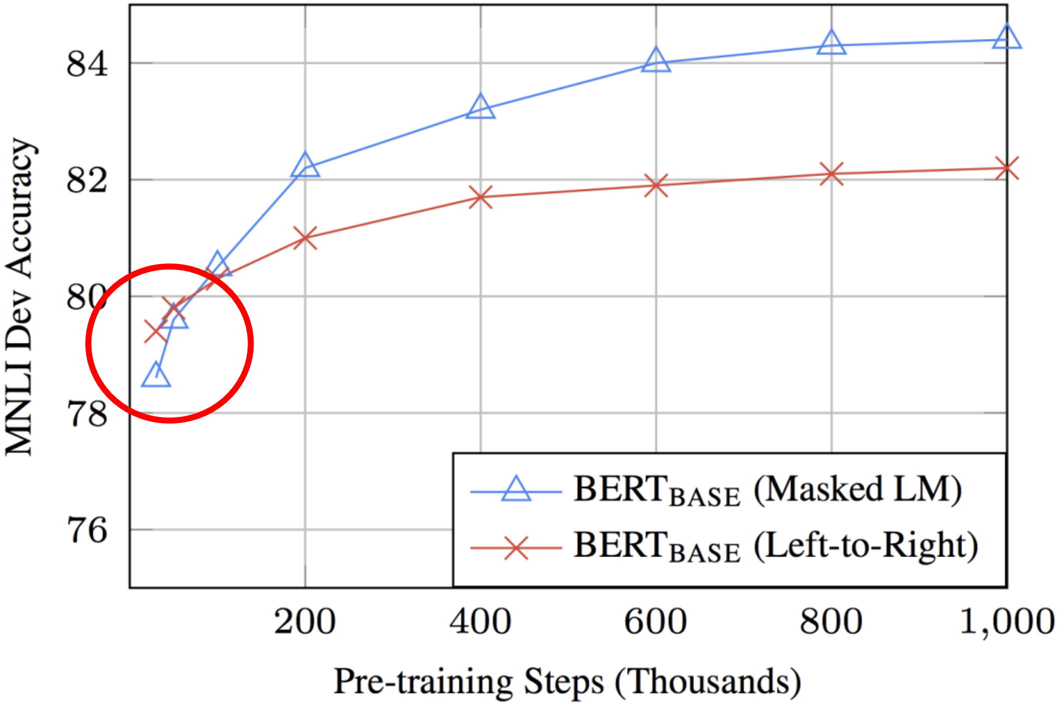


Masked Language Model Pretraining

- Bidirectional gives better performance

- But less efficient because only learn from 15% of tokens per example

- **Our method: best of both worlds**



New Pre-Training Task: Replaced Token Detection

- Instead of [MASK], replace tokens with plausible alternatives

the artist sold the painting

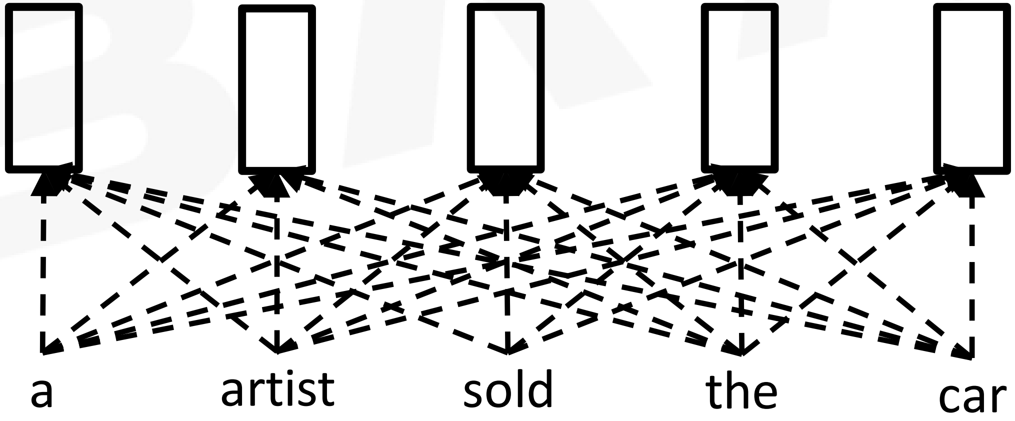
New Pre-Training Task: Replaced Token Detection

- Instead of [MASK], replace tokens with plausible alternatives

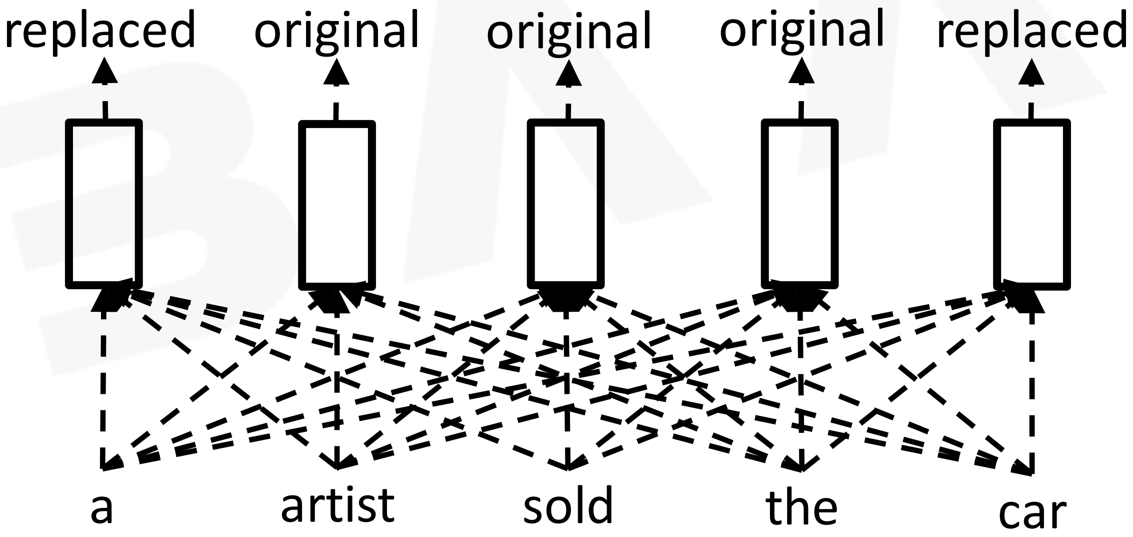
a car

~~the~~ artist sold the ~~painting~~

New Pre-Training Task: Replaced Token Detection

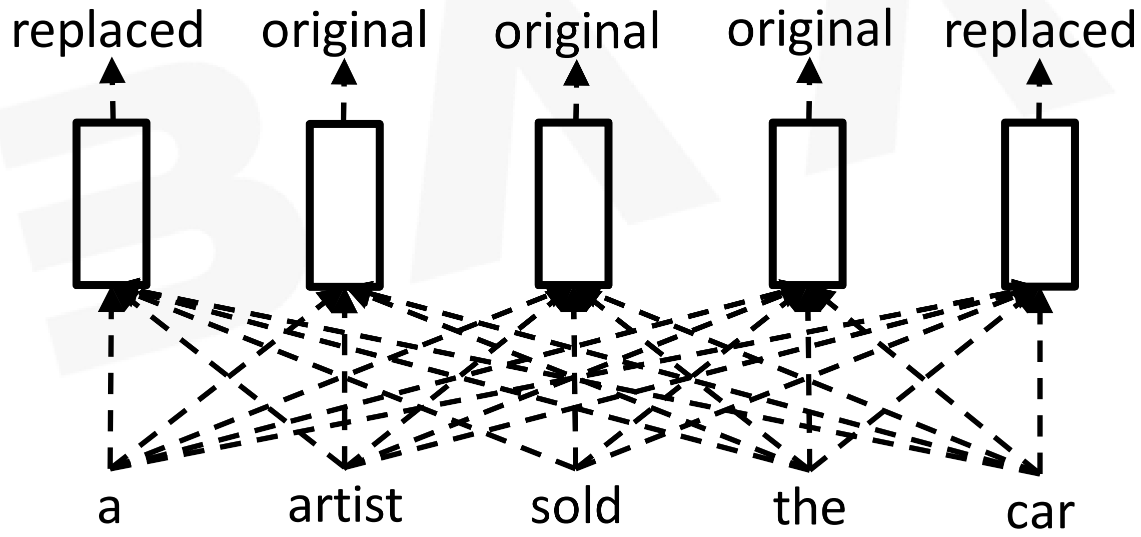


New Pre-Training Task: Replaced Token Detection



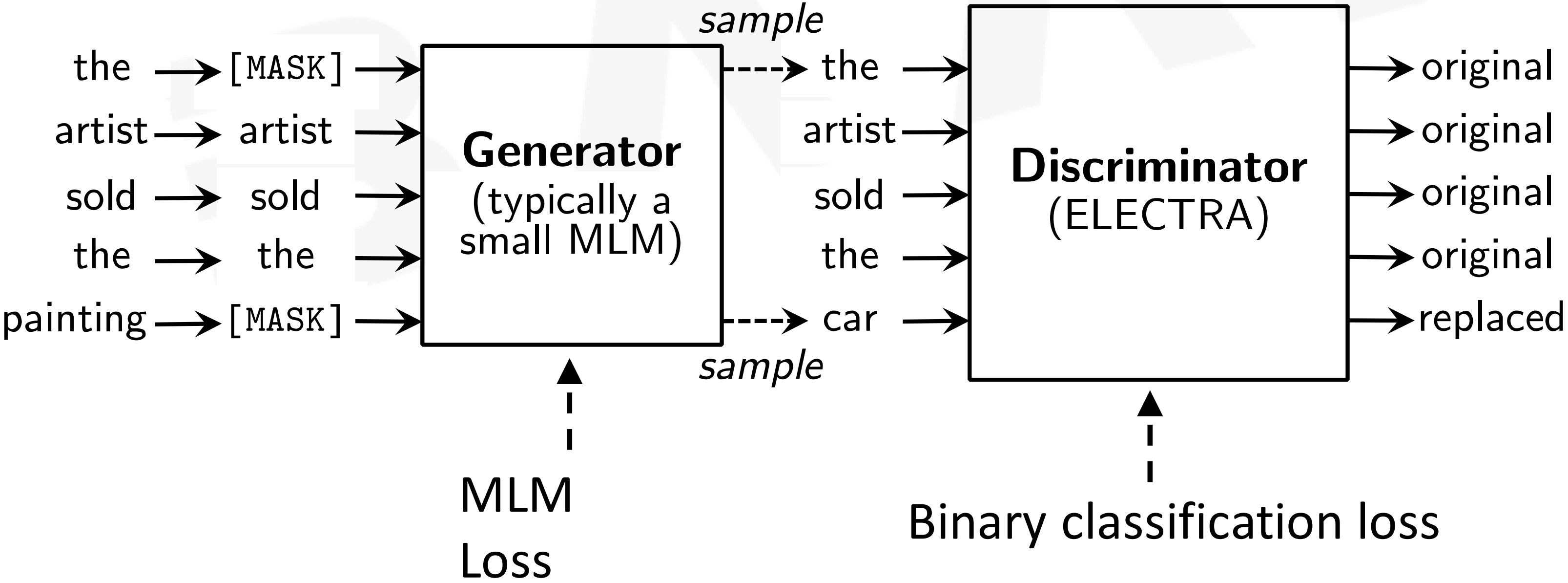
ELECTRA: "Efficiently Learning an Encoder to Classify Token Replacements Accurately"

Bidirectional model but learn from all tokens

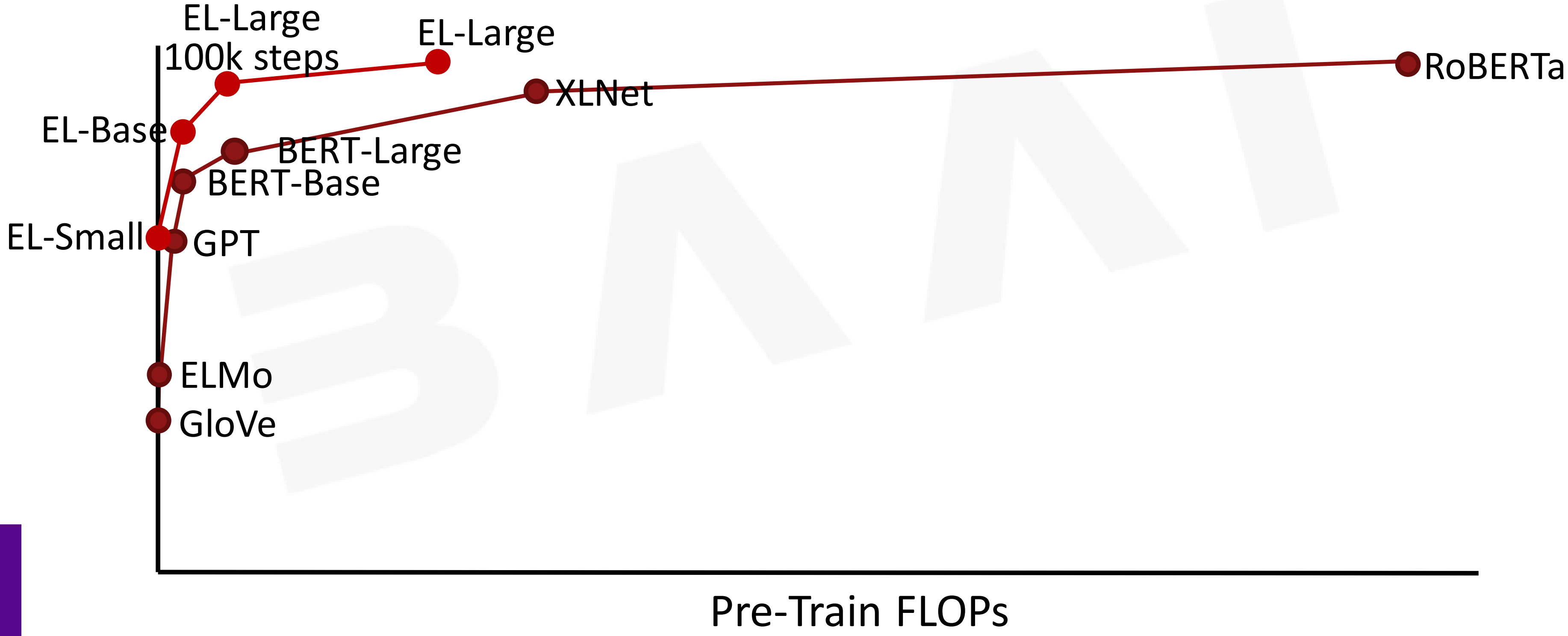


Generating Replacements

Plausible alternatives come from small masked language model (the “generator”) trained jointly with ELECTRA



Results: Glue Score vs Compute



Results: ELECTRA-Small

- Smaller model (1/3 hidden size) trained less (1/4 steps) as BERT-Base. Trains in 4 days on 1 V100 GPU.

Model	Train/Infer Speedup over BERT-Base	GLUE Score
ELMo	19x / 1.2x	71.2
GPT	1.6x / 1x	78.8
DistilBERT	- / 2x	77.8
BERT-Small (ours)	45x / 8x	74.7
ELECTRA-Small	45x / 8x	79.0
BERT-Base	1x / 1x	82.2

Results: ELECTRA-Large

- BERT-Large architecture, trained on XLNet data

Model	Train FLOPs	GLUE Score
BERT	0.3x	84.0
XLNet	1.3x	87.4
RoBERTa （100k steps）	0.9x	87.9
RoBERTa	4.5x	88.9
BERT-large (ours)	1x	87.2
ELECTRA	1x	89.0

Electra

- Recent pre-training methods let models benefit from unprecedented compute scale
    - But our environment/energy use doesn’t benefit!
    - It is important to be sensitive to compute when reporting results
- Replaced token detection is a more effective pre-training task then masked language modeling
    - Can provide good results on a single GPU in a few days
    - At larger scale, trains over 4x faster

Final thoughts

- Self-supervised (or “unsupervised”) learning is very successful for doing natural language understanding tasks
  - More so than conventional multi-task learning
  - There hasn’t (yet) been similar success for self-supervised learning in vision
- Has annotating lots of linguistic data all been a mistake?
  - Maybe. Language model learning exploits a much rich task compared to the categories in typical annotations

Final thoughts

- Is linguistic structure all a mistake?
  - No! Deep contextual word representations have phase-shifted from statistical association learners to **language discovery devices**!
  - Syntax emerges (approximately) in the geometry of BERT! See:
    - Kevin Clark, Urvashi Khandelwal, Omer Levy, & Christopher Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. BlackBoxNLP.
    - John Hewitt and Christopher Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. NAACL.
- Does going big stretch any analogy to child language acquisition?
  - Maybe, but it's more that acquisition without grounding is unrealistic

Recent Breakthroughs in Natural Language Processing

Stanford

Christopher Manning

Director, Stanford Artificial Intelligence Laboratory

@chrmanning ❀ @stanfordnlp

BAAI 2019

BAAI CONFERENCE

2019北京智源大会

自然语言处理专题论坛

微信扫码加入

智源大会现场微信交流群