Lecture 20 The Future of NLP + Deep Learning

Deep Learning for NLP 5 years ago

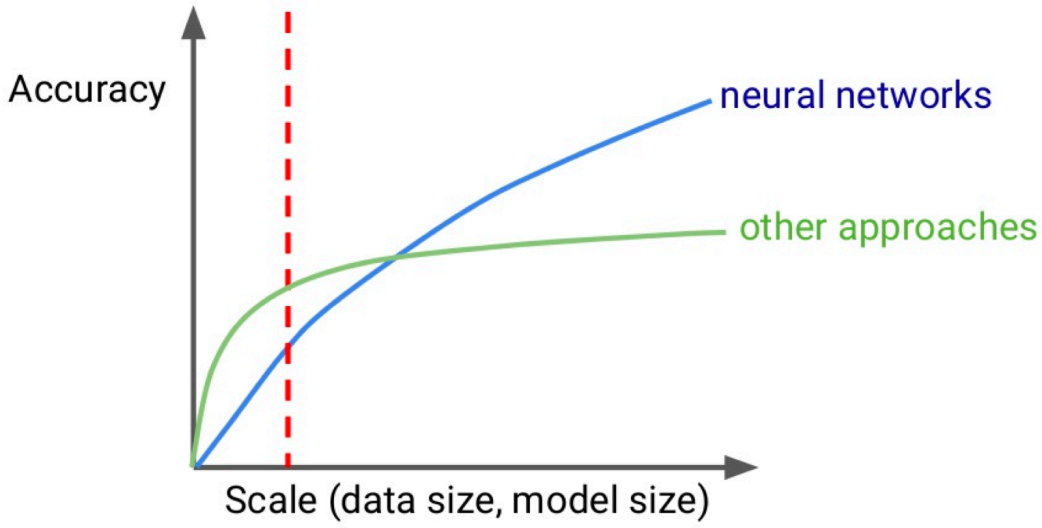
- No Seq2Seq
- No Attention
- No large-scale QA/reading comprehension datasets
- No TensorFlow or Pytorch

Future of Deep Learning + NLP

- 利用无标签数据
  - Back-translation 和 无监督机器翻译
  - 提高预训练和GPT-2
- 接下来呢?
  - NLP技术的风险和社会影响
  - 未来的研究方向

Why has deep learning been so successful recently?

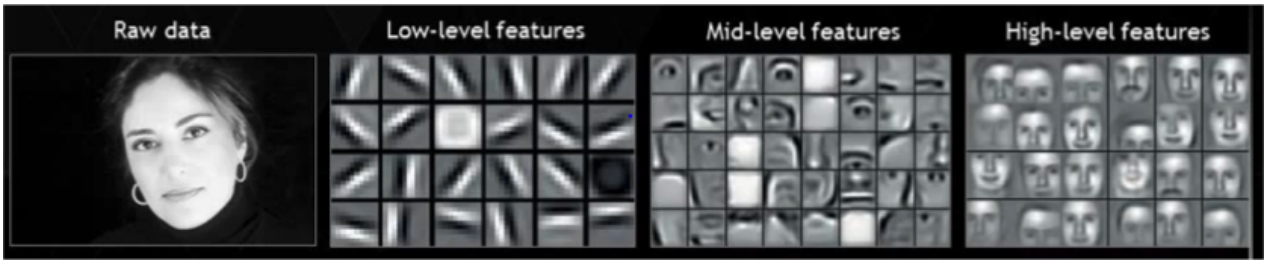
1980s and 1990s



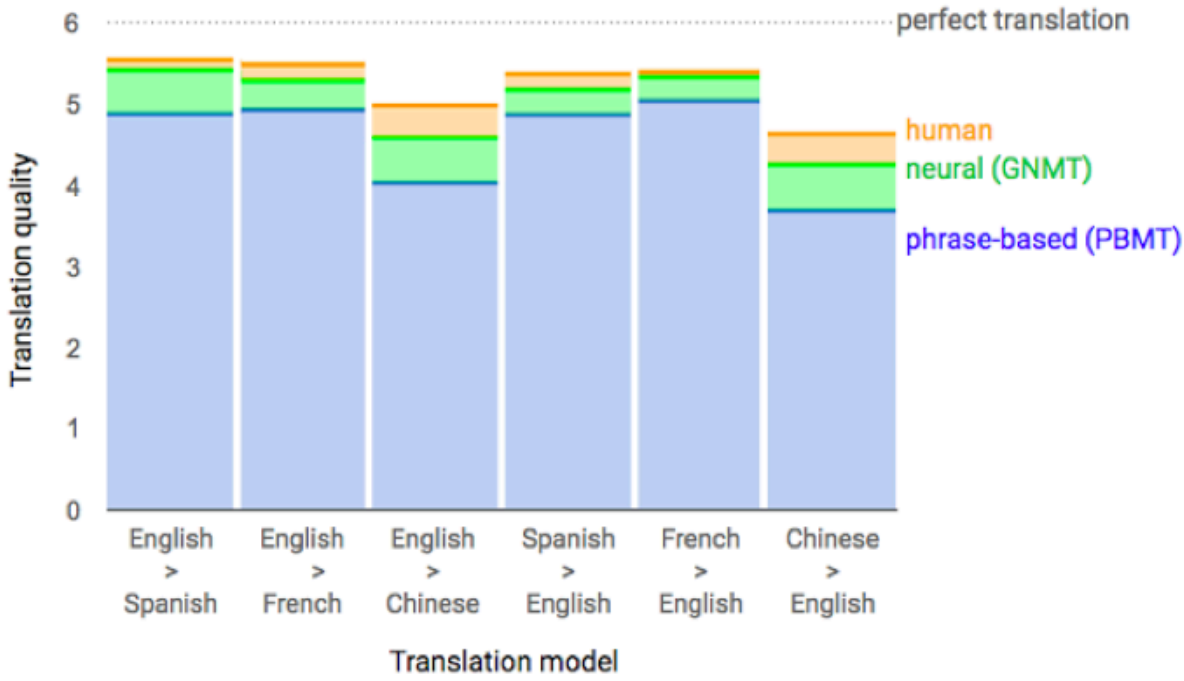
- 扩展能力（模型和数据大小）是深度学习近些年来成功的原因
- 过去受到计算资源和数据资源的规模限制

Big deep learning successes

三个使用大量数据获得成功的范例



- 图像识别：被 Google, Facebook 等广泛使用
  - ImageNet: 14 million examples



- 机器翻译：谷歌翻译等
  - WMT: Millions of sentence pairs

- 打游戏：Atari Games, AlphaGo, and more
  - 10s of millions of frames for Atari AI
  - 10s of millions of self-play games for AlphaZero

NLP Datasets

- 即使是英语，大部分任务也只有 100k 或更少的有标签样本
- 其他语言的可用数据就更少了
  - 有成千上万的语言，其中有成百上千的语言的母语使用者是大于一百万的
  - 只有 10% 的人将英语作为他们的第一语言
- 越来越多的解决方案是使用 **无标签** 数据

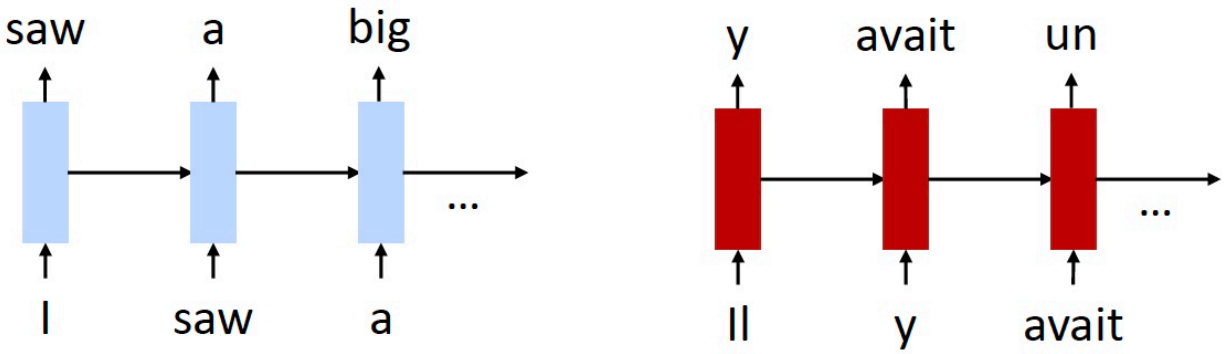
Using Unlabeled Data for Translation

Machine Translation Data

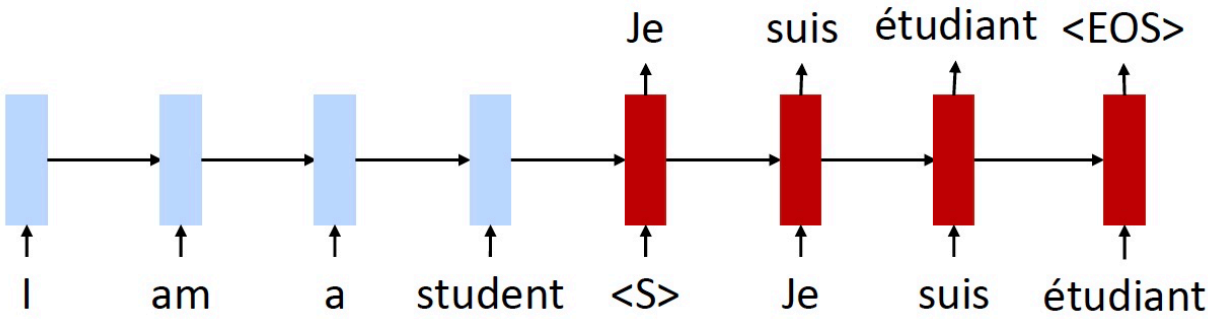
- 获得翻译需要人类的专业知识
  - 限制数据的大小和领域

- 语言文本更容易获得

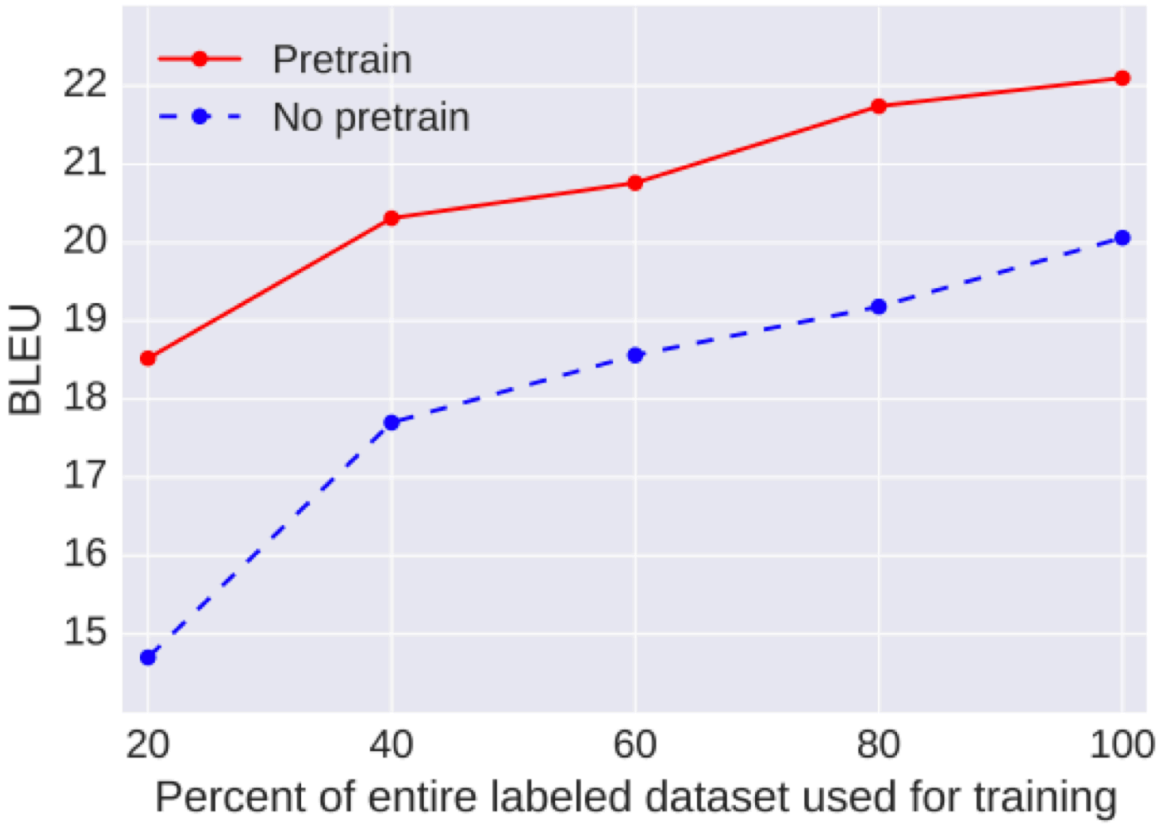
Pre-Training



- 分别将两个预训练好的语言模型作为 Encoder 和 Decoder



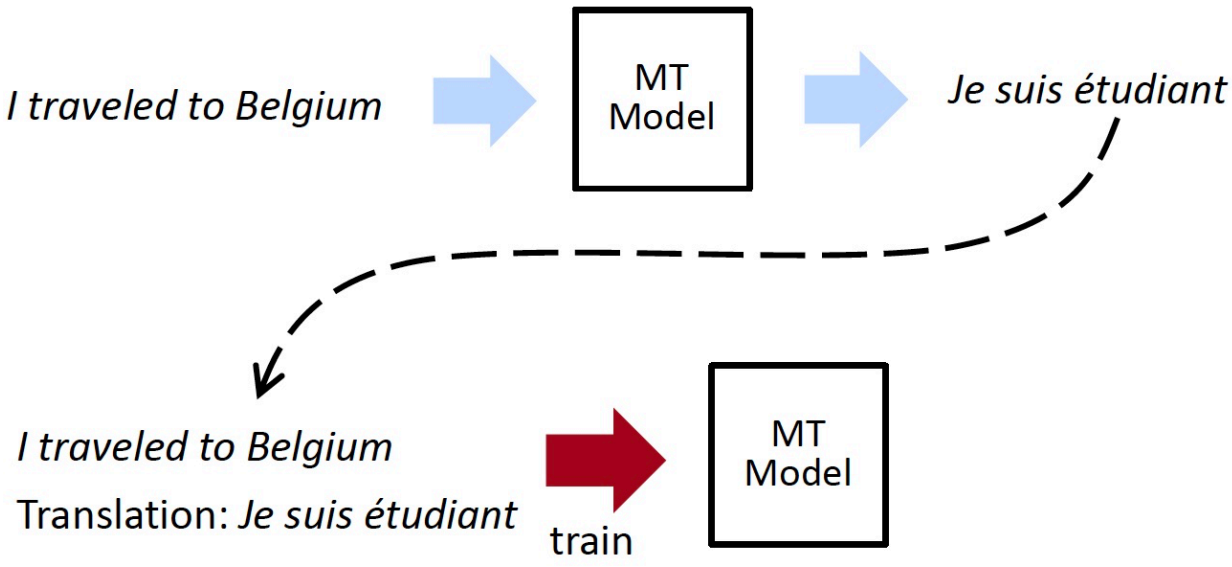
- 然后使用双语数据共同训练



- English -> German Results: 2+ BLEU point improvement

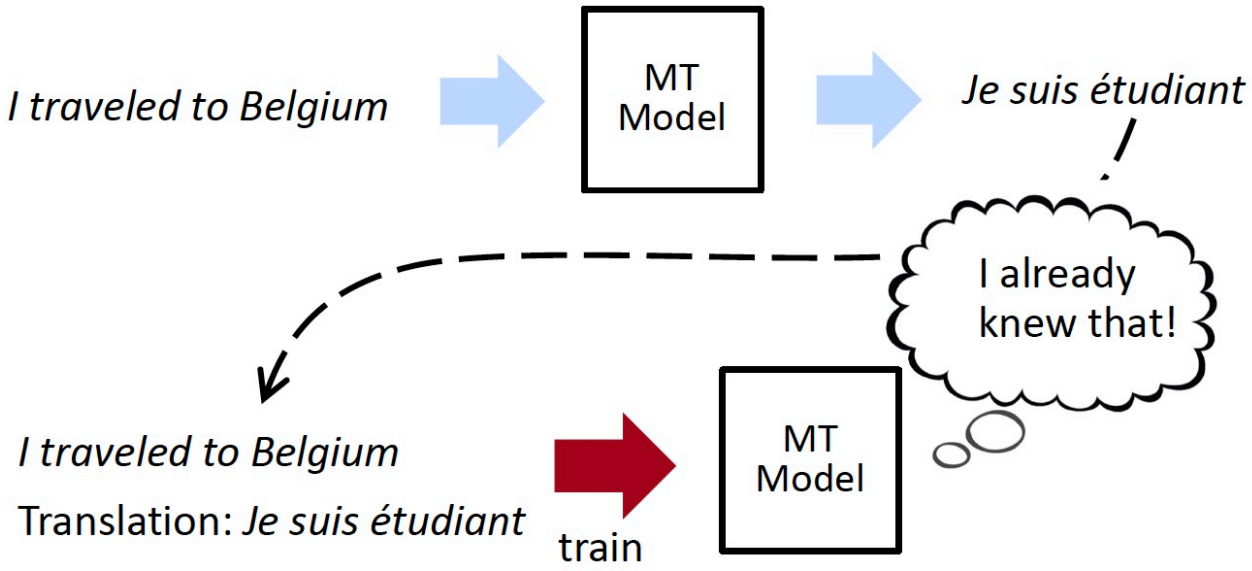
Ramachandran et al., 2017

Self-Training



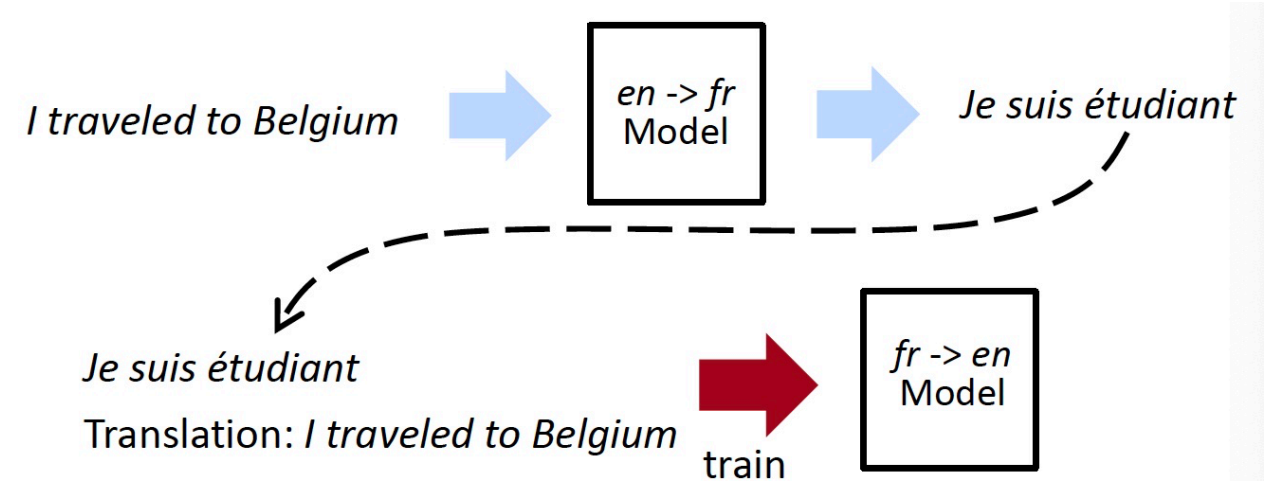
- 预训练的问题：预训练中两个语言之间没有交互
- 自训练：标记未标记的数据以获得有噪声的训练样本

- Circular?



- 自训练技术没有被广泛使用，因为其训练的来源是其之前的产出

Back-Translation



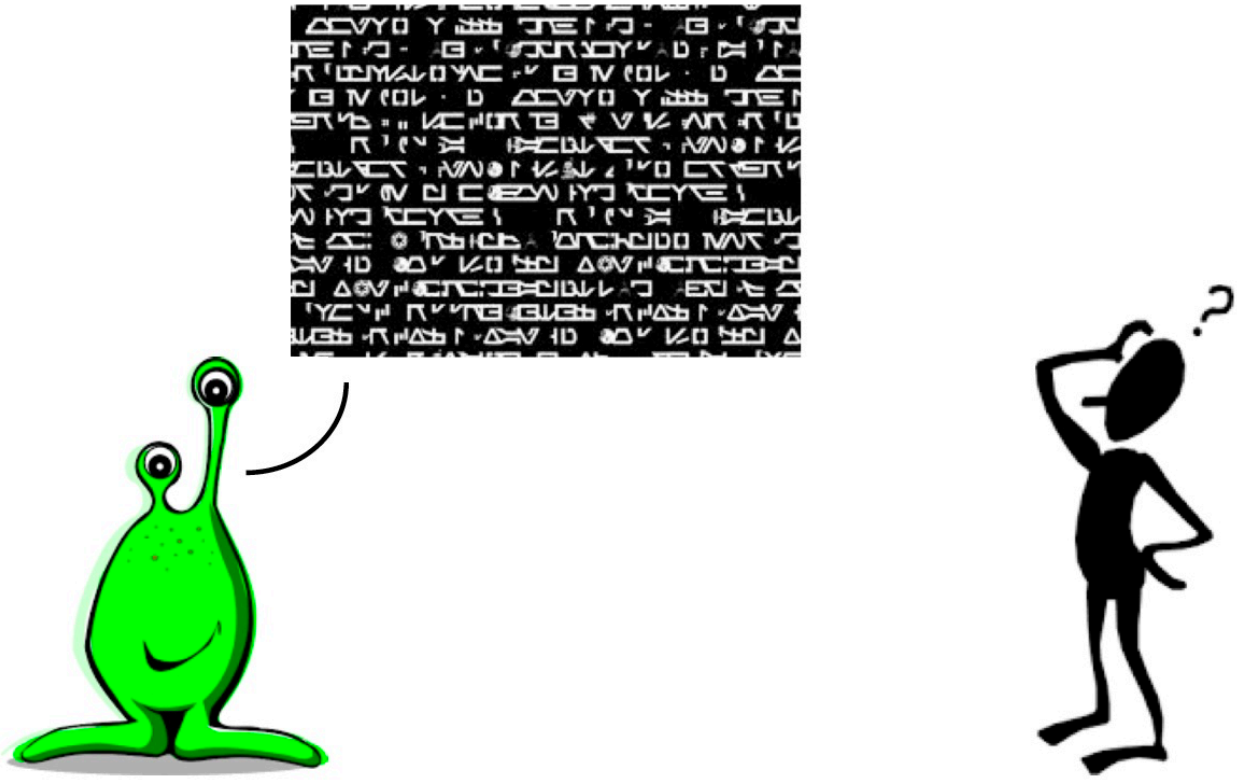
- 有两种方向相反的机器翻译模型 $\text{en} \to \text{fr}$ 和 $\text{fr} \to \text{en}$
- 不再循环
- 模型再也看不到“坏”翻译,只有坏输入
- 模型训练时会加入一些标记数据，确保 $\text{en} \to \text{fr}$ 模型的输出，即 $\text{fr} \to \text{en}$ 模型的输入，从而保证模型的正常
- 如何协调对标记数据与未标记数据的训练呢?
  - 先在标记数据上训练两个模型
  - 然后在未标记数据上标记一些数据
  - 再在未标记数据上进行反向翻译的训练
  - 重复如上的过程

Large-Scale Back-Translation

Citation	Model	BLEU
Shazeer et al., 2017	Best Pre-Transformer Result	26.0
Vaswani et al., 2017	Transformer	28.4
Shaw et al, 2018	Transformer + Improved Positional Embeddings	29.1
Edunov et al., 2018	**Transformer + Back-Translation**	**35.0**

- 4.5M English-German sentence pairs and 226M monolingual sentences

What if there is no Bilingual Data?



当我们只有未标记的句子时，我们使用一种比完全的翻译更简单的任务

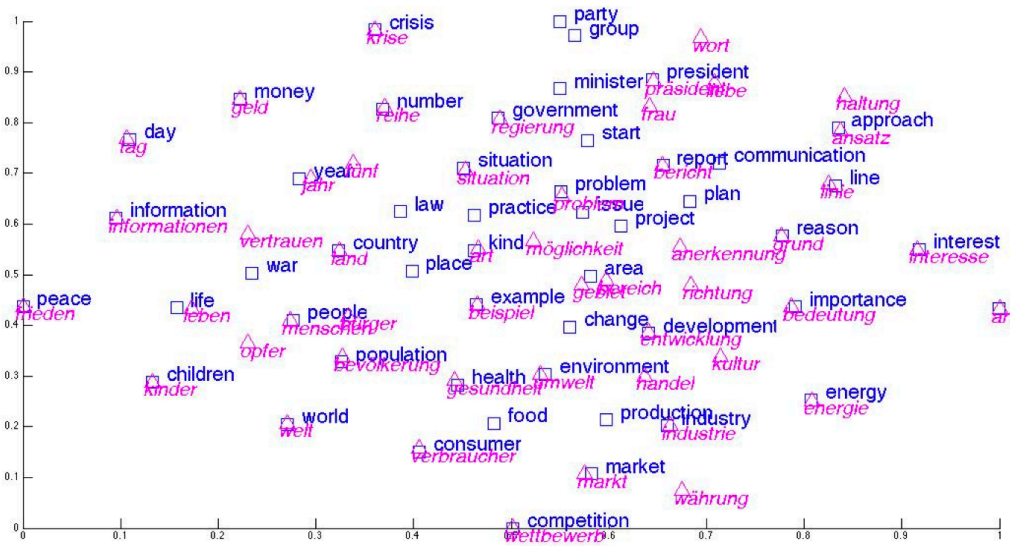
- 不是做句子翻译
- 而是做单词翻译

我们想要找到某种语言的翻译但不使用任何标记数据

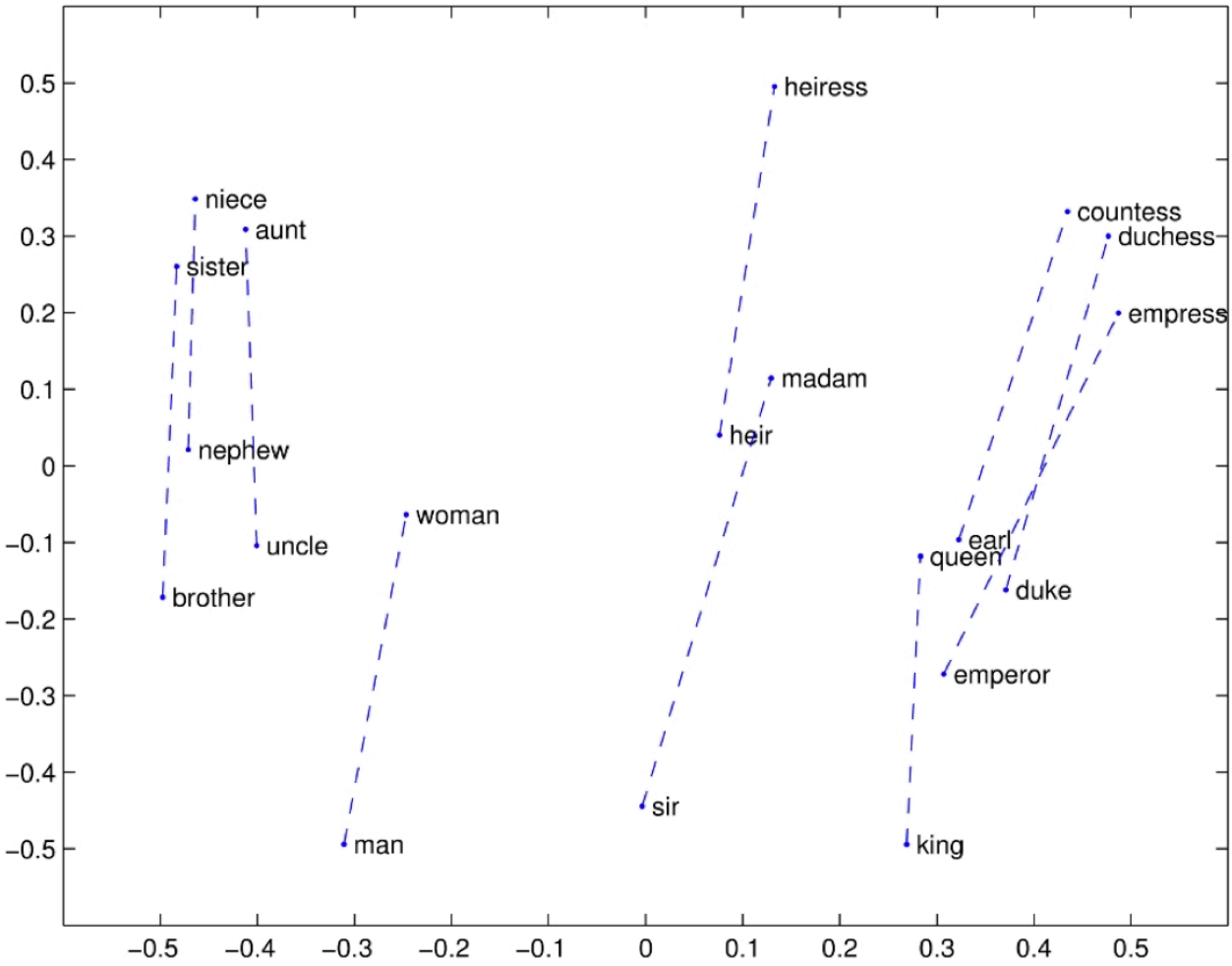
Unsupervised Word Translation

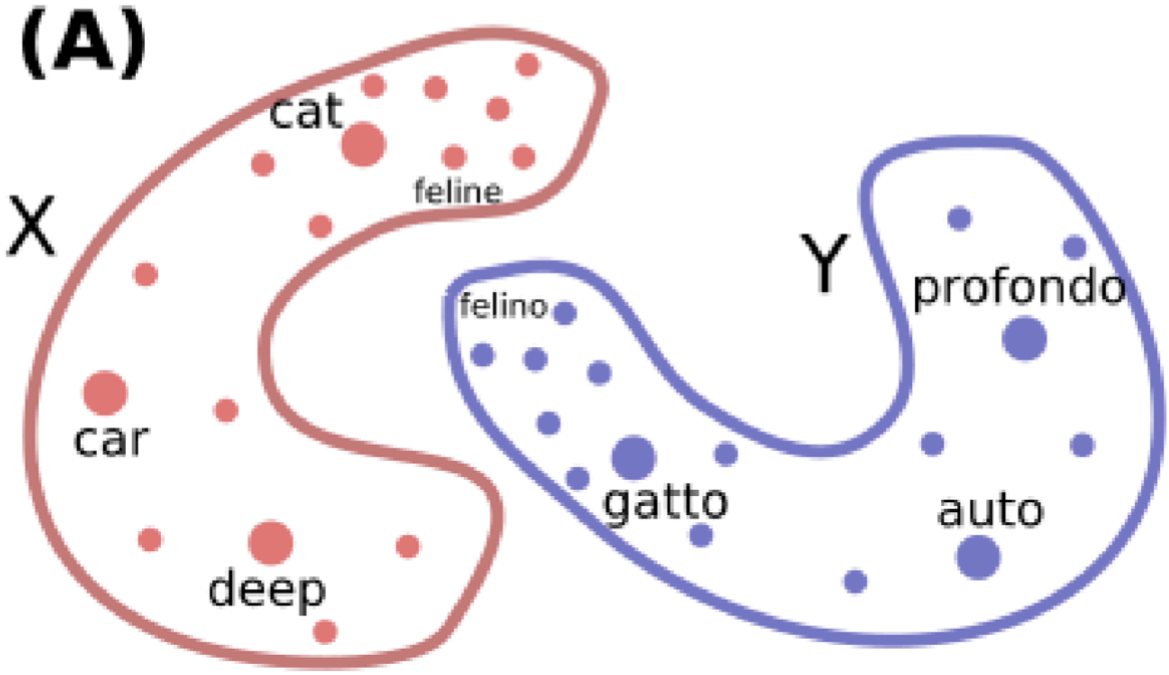
- 跨语言文字嵌入 cross-lingual word embeddings
  - 两种语言共享嵌入空间
  - 保持词嵌入的正常的好属性
  - 但也要接近他们的翻译

- 想从单语语料库中学习



- 如上图所示，在共享的嵌入空间中，每个英文单词都有其对应的德语单词，并且距离很近
- 我们在使用时，只需选取英文单词在嵌入空间中距离最近的德语单词，就可以获得对应的翻译

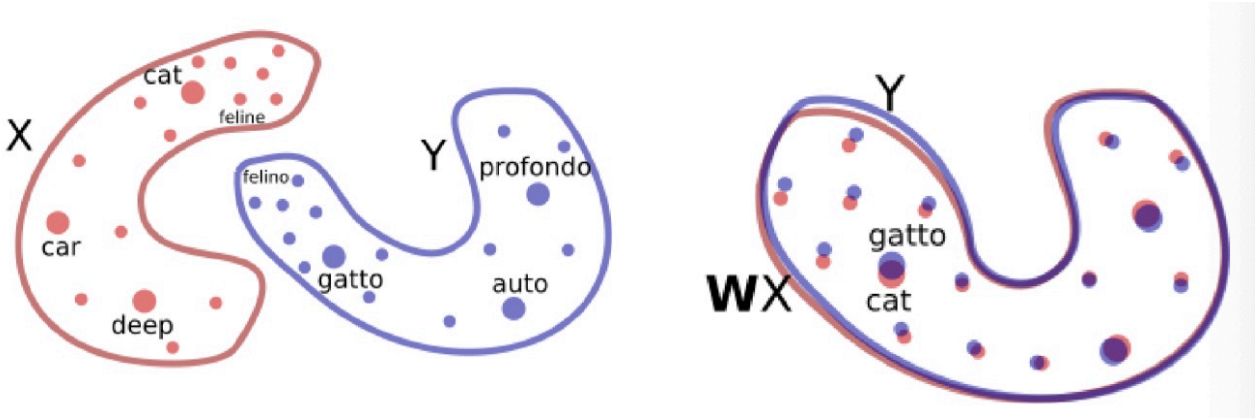




- 词嵌入有很多结构
- 假设:不同语言之间的结构应该相似

即使是运行两次 word2vec 会获得不同的词嵌入，嵌入空间的结构有很多规律性

- 如上图所示，是英语与意大利语的词嵌入，矢量空间看上去彼此十分不同，但是结构是十分相似的
    - 可以理解为，在英语词嵌入空间中的 cat 与 feline 的距离与意大利语词典如空间中的 gatto 和 felino 之间的距离是相似的
- 我们在跨语言的词嵌入中想要学习不同种语言的词嵌入之间的对齐方式

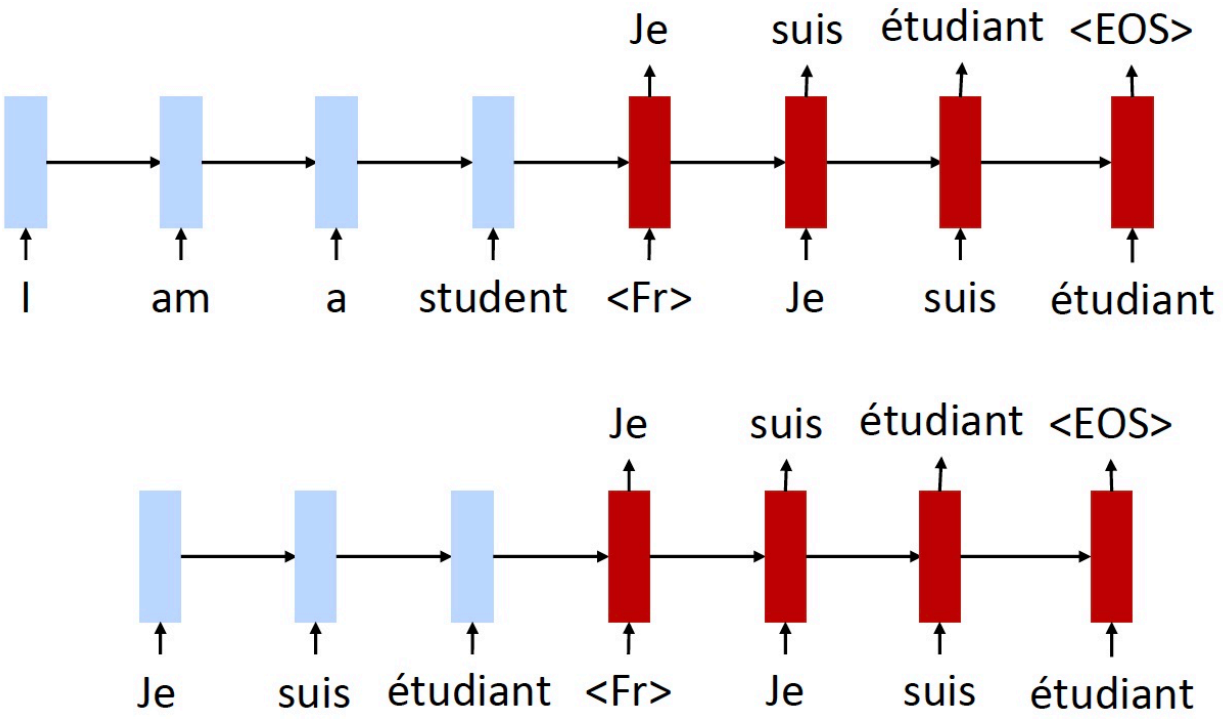


- 首先在单语语料库上运行 word2vec 以得到单词嵌入 X 和 Y
- 学习一个（正交）矩阵 W 使得 WX~Y
    - 使用对抗训练来学习 W
    - 鉴别器：预测一个嵌入是来自于 Y 的还是来自于 X 并使用 W 转换后的嵌入
    - 训练 W 使得鉴别器难以区分这两者
    - 其他可以被用来进一步提升效果的方法参见 [Word Translation without Parallel Data]
    - 正交性来约束词嵌入的原因是为了防止过拟合
        - 我们假设我们的嵌入空间是类似的，只是需要对英语的词向量和意大利语的词向量进行

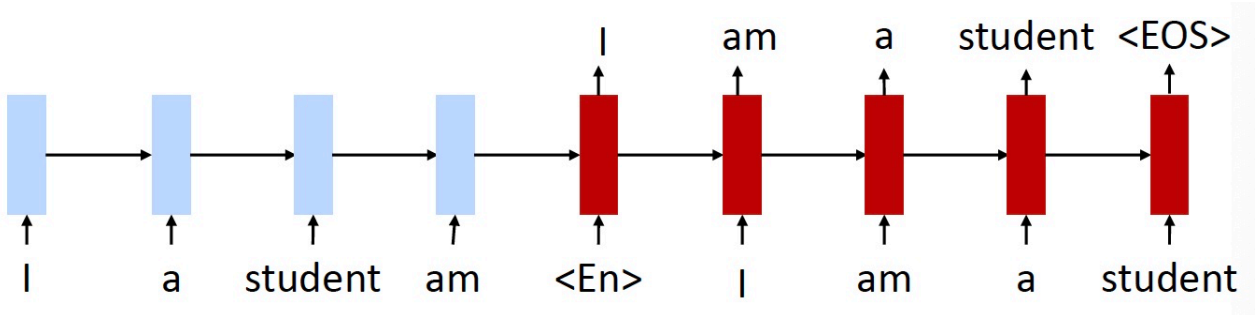
旋转

Unsupervised Machine Translation

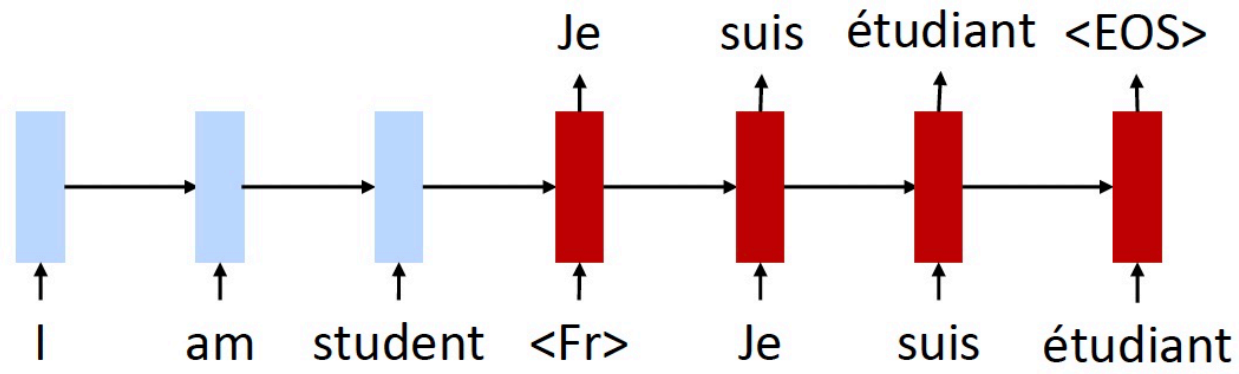
- 模型：不考虑不同输入和输出语言，使用相同的(共享的) encoder-decoder (没有使用注意力)
  - 使用 cross-lingual 的词嵌入来初始化，即其中的英语和法语单词应该看起来完全相同
- 我们可以喂给 encoder 一个英文句子，也可以喂一个法语句子，从而获得 cross-lingual embeddings，即英文句子和法语句子中各个单词的词嵌入，这意味着 encoder 可以处理任何输入
- 对于 decoder，我们需要喂一个特殊的标记 $<\mathrm{Fr}>$ 来告诉模型应该生成什么语言的输出
  - 可以用做一个 auto-encoder，完成 $en \rightarrow en$，即再现输入序列



接下里是模型的训练过程

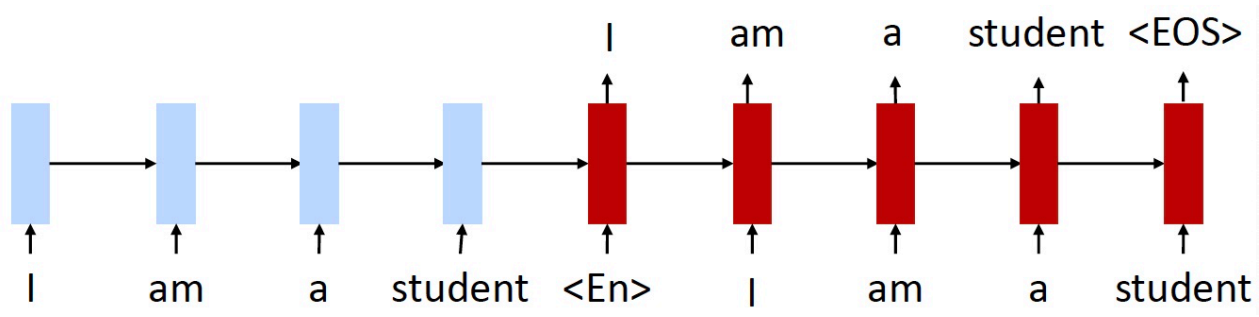


- Training objective 1: de-noising autoencoder 去噪自编码器
  - 输入一个打乱后的英语 / 法语句子
  - 输出其原来的句子
  - 由于这是一个没有注意力机制的模型，编码器将整个源句子转换为单个向量，自编码器的作用是保证来自于 encoder 的向量包含和这个句子有关的，能使得我们恢复原来的句子的所有信息

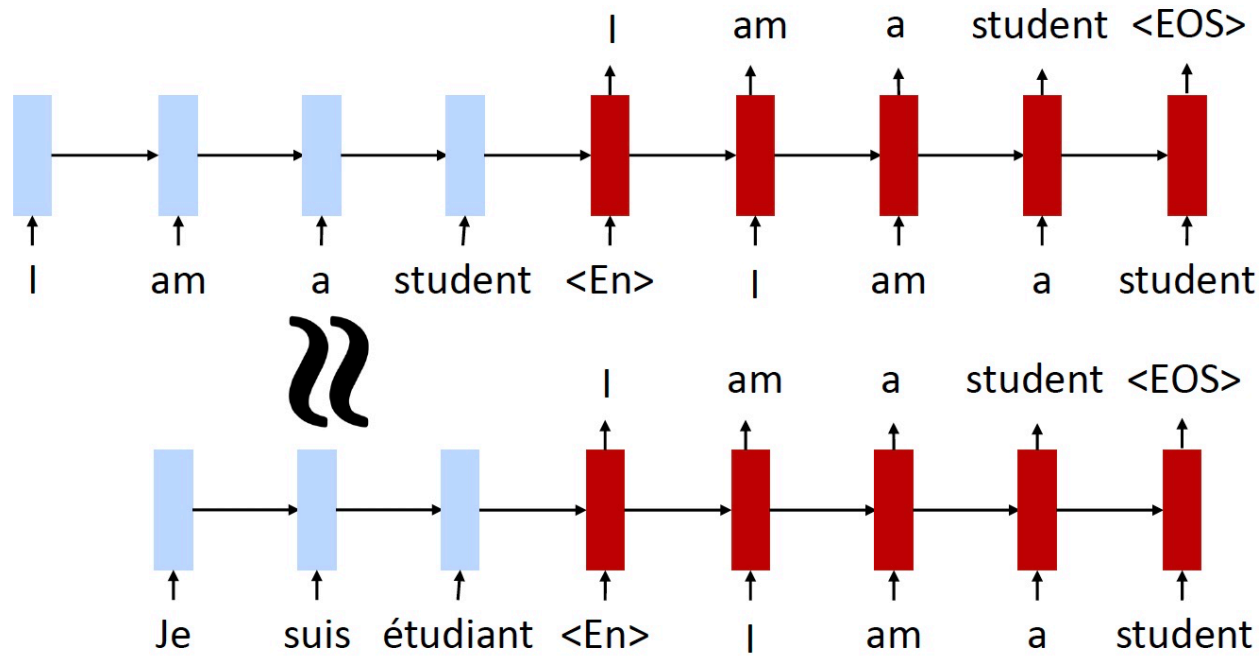


- Training objective 2: back translation (只有无标签的数据)
  - 首先翻译 $fr \rightarrow en$
  - 然后使用一个监督样本来训练 $en \rightarrow fr$
- 注意，这里的 $fr \rightarrow en$ 输出的句子，是 $en \rightarrow fr$ 输入的句子，这个句子是有些混乱的，不完美的，例如这里的 “I am student”，丢失了 “a”
- 我们需要训练模型，即使是有这样糟糕的输入，也能够还原出原始的法语句子

Why Does This Work?

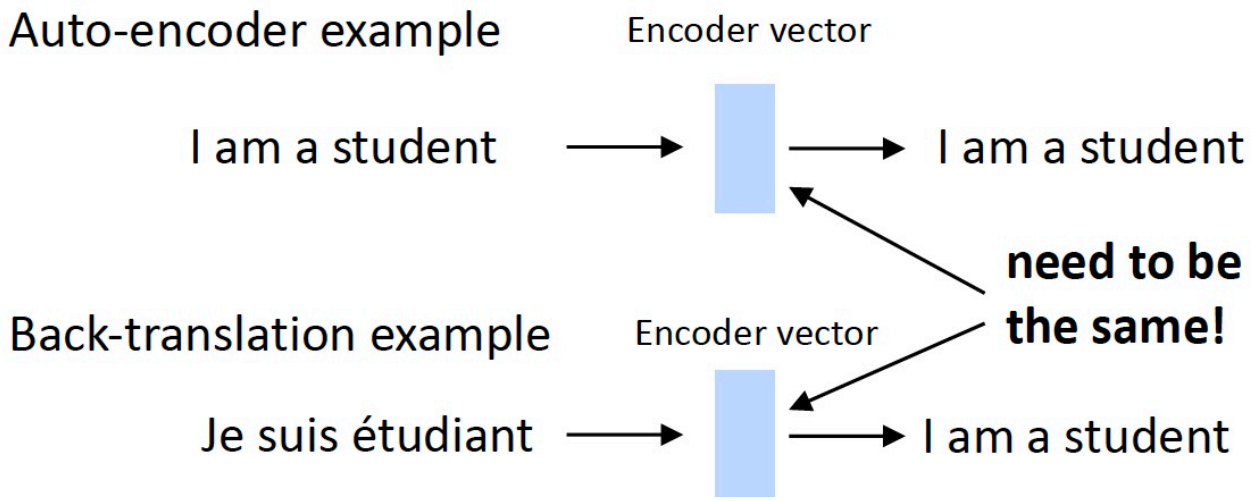


- 跨语言嵌入和共享编码器为模型提供了一个起点
  - 使用 cross-lingual 的词嵌入来初始化，即其中的英语和法语单词应该看起来完全相同



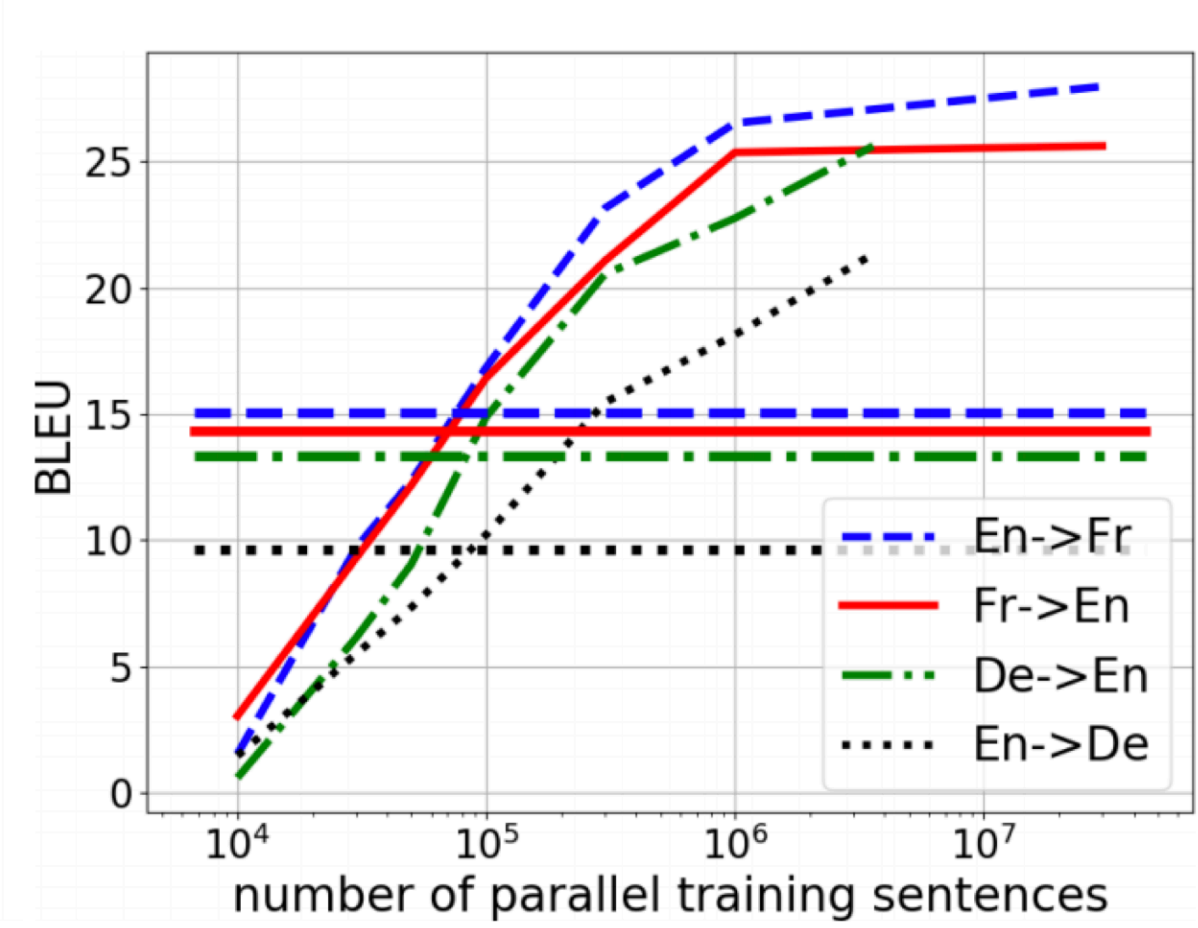
- 共享编码器

- 例如我们以一个法语句子作为模型的输入
- 由于嵌入看起来非常相似，并且我们使用的是相同的 encoder
- 因此 encoder 得到的法语句子的 representation 应该和英语句子的 representation 非常相似
- 所以希望能够获得和原始的英语句子相同的输出



- 目标鼓励 language-agnostic 语言无关的表示
  - 获得一个与语言类型无关的 encoder vector

Unsupervised Machine Translation



> Lample et al., 2018

- 水平线是无监督模型，其余的都是有监督的
- 在一定的监督数据规模下，无监督模型能够取得和监督模型类似的效果
- 当然，随着数据规模的增大，监督模型的效果会提升，超过无监督模型

Attribute Transfer

还可以使用无监督的机器翻译模型完成属性转移

Relaxed ↔ Annoyed	
Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night 💗🎄✨
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend 😡😡😡
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 😒
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 😊
Male ↔ Female	
Male	Gotta say that beard makes you look like a Viking...
Female	Gotta say that hair makes you look like a Mermaid...
Female	Awww he's so gorgeous 😍 can't wait for a cuddle. Well done 😘 xxx
Male	Bro he's so f***ing dope can't wait for a cuddle. Well done bro

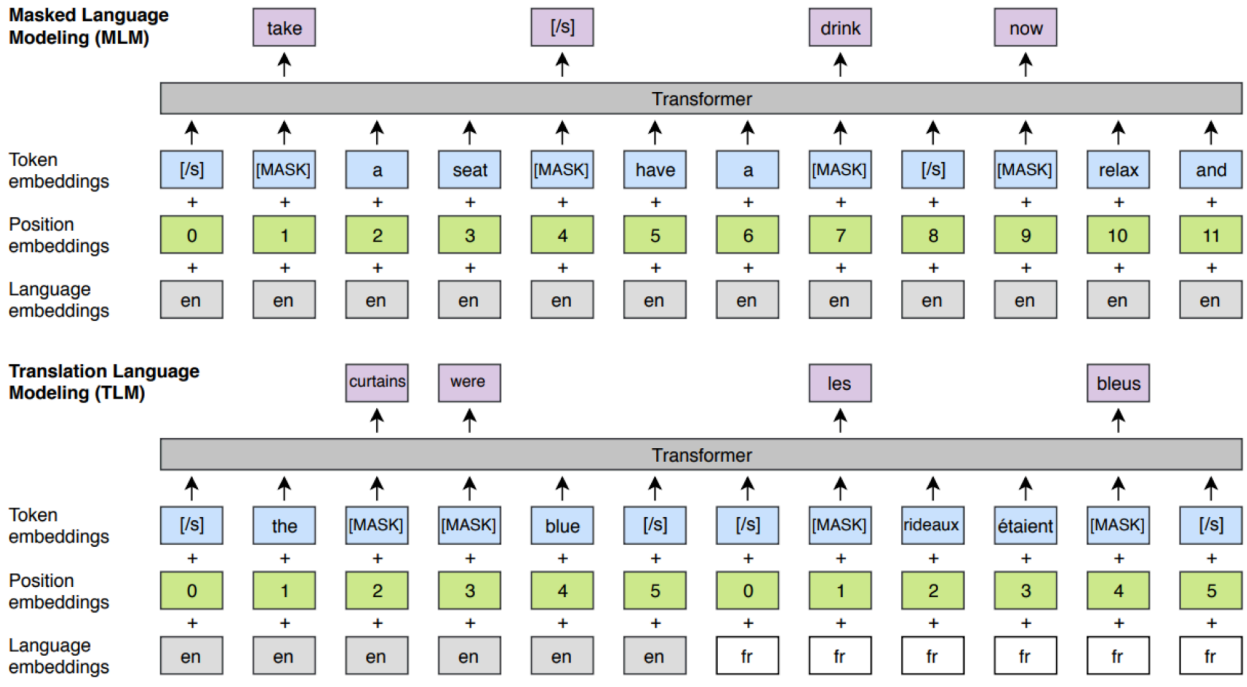
> Lample et al., 2019

- Collector corpora of "relaxed" and "annoyed" tweets using hashtags
- Learn un unsupervised MT model

Not so Fast

- 英语，法语和德语是相当类似的语言
- 在非常不同的语言上（例如英语和土耳其语）
  - 完全的无监督的词翻译并不十分有效。需要种子字典可能的翻译
    - 简单的技巧：使用相同的字符串从词汇
  - UNMT几乎不工作

Cross-Lingual BERT



> Lample and Conneau., 2019

- 上图 1 是常规的 BERT，有一系列的英语句子，并且会 mask 一部分单词
  - 谷歌实际上已经完成的是训练好的多语言的 BERT
  - 基本上是连接一大堆不同语言的语料库，然后训练一个模型
    - masked LM training objective
- 上图 2 是 Facebook 提出的
  - 联合了 masked LM training objective 和 翻译
  - 给定一个英语句子和法语句子，并分别 mask 一部分单词，并期望模型填补

Unsupervised MT Results

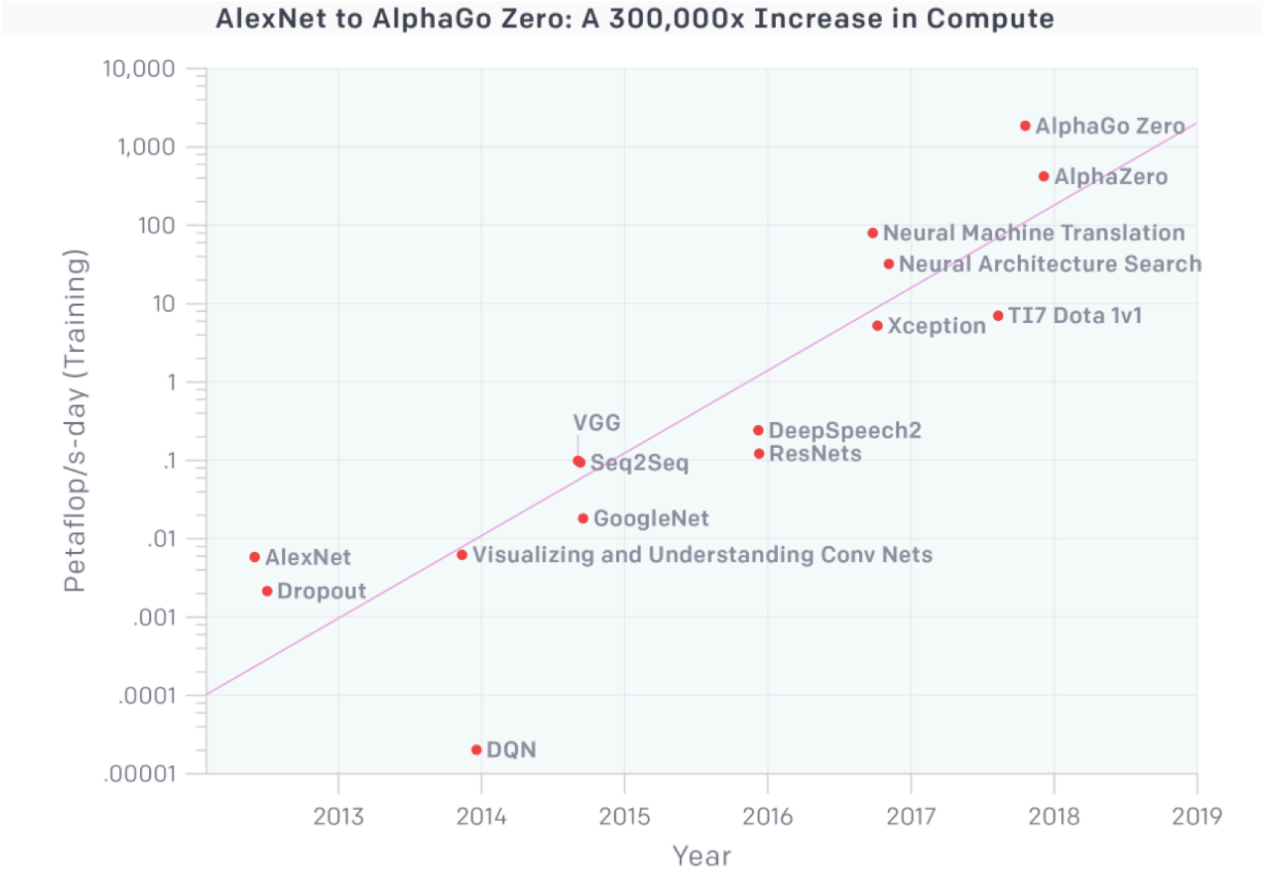
Model	En-Fr	En-De	En-Ro
UNMT	25.1	17.2	21.2
UNMT + Pre-Training	33.4	26.4	**33.3**
Current supervised State-of-the-art	**45.6**	**34.2**	29.9

Huge Models and GPT-2

Training Huge Models

Model	# Parameters
Medium-sized LSTM	10M
ELMo	90M
GPT	110M
BERT-Large	320M
GPT-2	1.5B
Honey Bee Brain	~1B synapses

This is a General Trend in ML



- peta : 用于计量单位，表示10的15次方，表示千万亿次
- FLOPS = FLoating-point Operations Per Second，每秒浮点运算次数

Huge Models in Computer Vision

LARGE SCALE GAN TRAINING FOR
HIGH FIDELITY NATURAL IMAGE SYNTHESIS

Andrew Brock*† Heriot-Watt University ajb5@hw.ac.uk
Jeff Donahue† DeepMind jeffdonahue@google.com
Karen Simonyan† DeepMind simonyan@google.com

- 150M parameters

GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism

Yanping Huang
Google Brain
huangyp@google.com

Youlong Cheng
Google Brain
ylc@google.com

Dehao Chen
Google Brain
dehao@google.com

HyoukJoong Lee
Google Brain
hyouklee@google.com

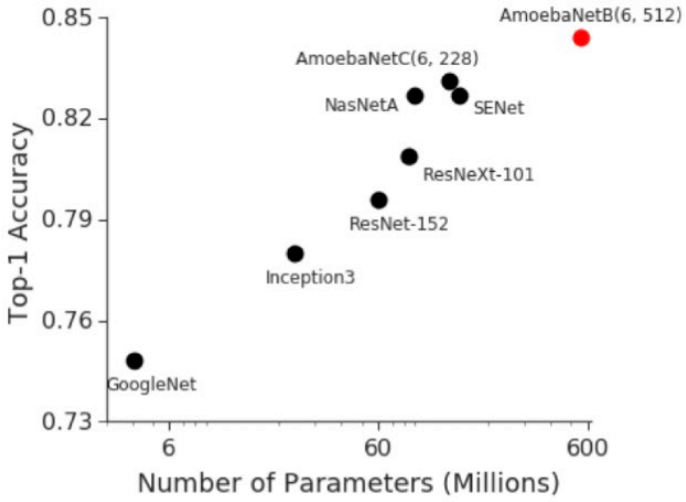
Jiquan Ngiam
Google Brain
jngiam@google.com

Quoc V. Le
Google Brain
qvl@google.com

Zhifeng Chen
Google Brain
zhifengc@google.com

- 550M parameters

ImageNet Results

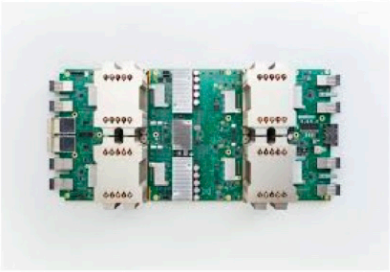


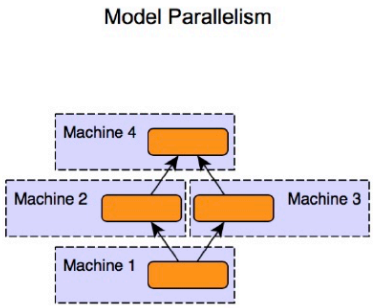
Training Huge Models

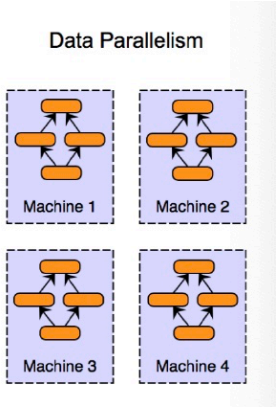
Mesh-TensorFlow: Deep Learning for Supercomputers

Noam Shazeer, Youlong Cheng, Niki Parmar,
Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee
Mingsheng Hong, Cliff Young, Ryan Sepassi, Blake Hechtman
Google Brain
{noam, ylc, nikip, trandustin, avaswani, penporn, phawkins,
hyouklee, hongm, cliffy, rsepassi, blakehechtman}@google.com

- 更好的硬件
- 数据和模型的并行化







GPT-2

- 只是一个非常大的 Transformer LM
- 40 GB的训练文本
  - 投入相当多的努力去确保数据质量
  - 使用 reddit 中获得高投票的网页 link

So What Can GPT-2 Do?

- 显然，语言建模(但是非常好)
- 数据集上得到最先进的困惑，甚至没有训练

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

> Radford et al., 2019

- **Zero-Shot Learning**: no supervised training data! 在没有接受过训练的情况下常识完成任务
  - Ask LM to generate from a prompt
- **Reading Comprehension:** <context> <question> A:
- **Summarization:** <article> TL;DR:
- **Translation:**

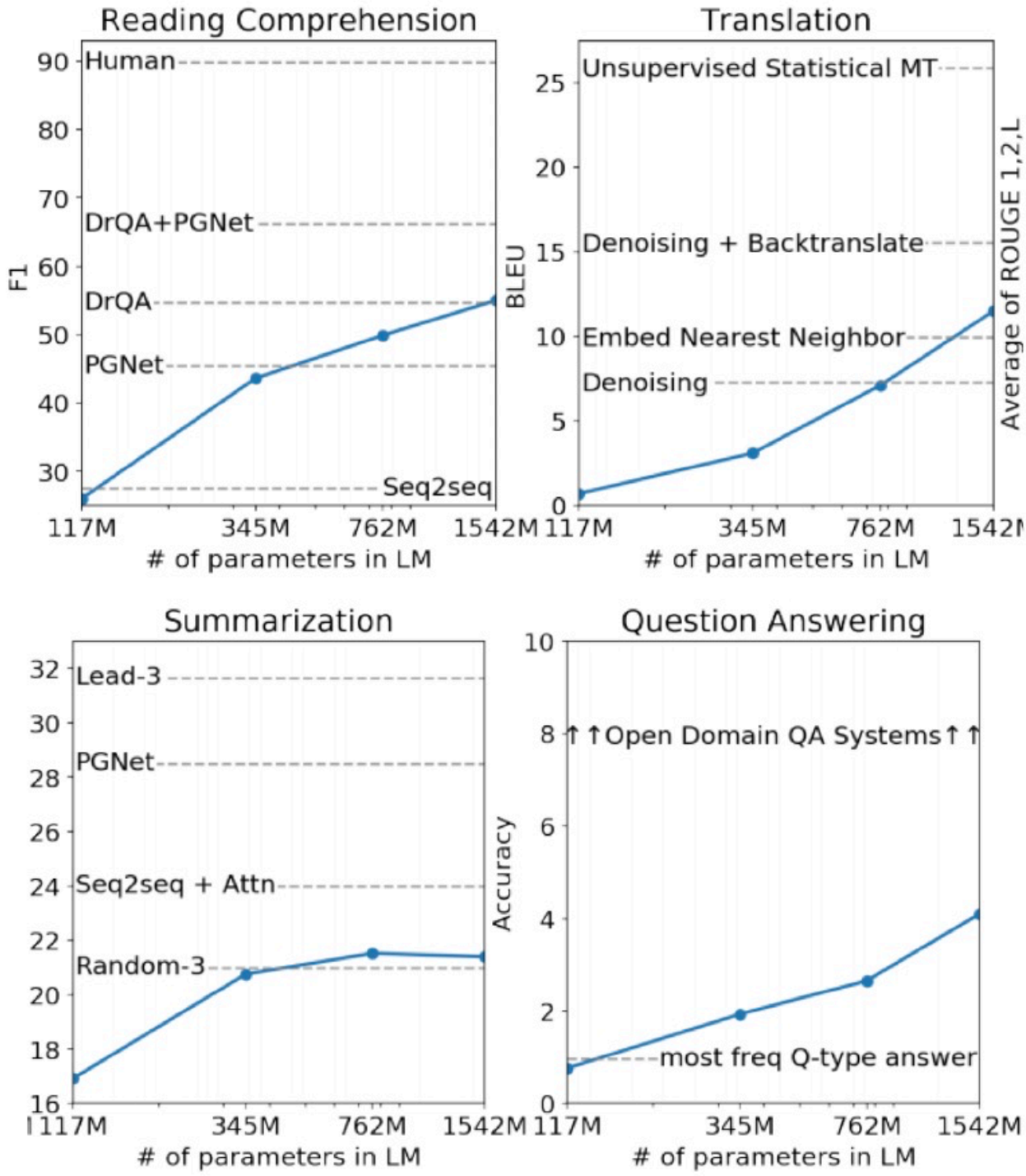
  <English sentence1> = <French sentence1>

  <English sentence 2> = <French sentence 2>

  .....

  <Source sentenc> =
- **Question Answering:** <question> A:

GPT-2 Results



How can GPT-2 be doing translation?

- 它有一个很大的语料库，里面几乎全是英语

”I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool].**

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”**Mentez mentez, il en restera toujours quelque chose,**” which translates as, ”**Lie lie and something will always remain.**”

“I hate the word ‘**perfume,**”’ Burr says. ‘It’s somewhat better in French: ‘**parfum.**’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

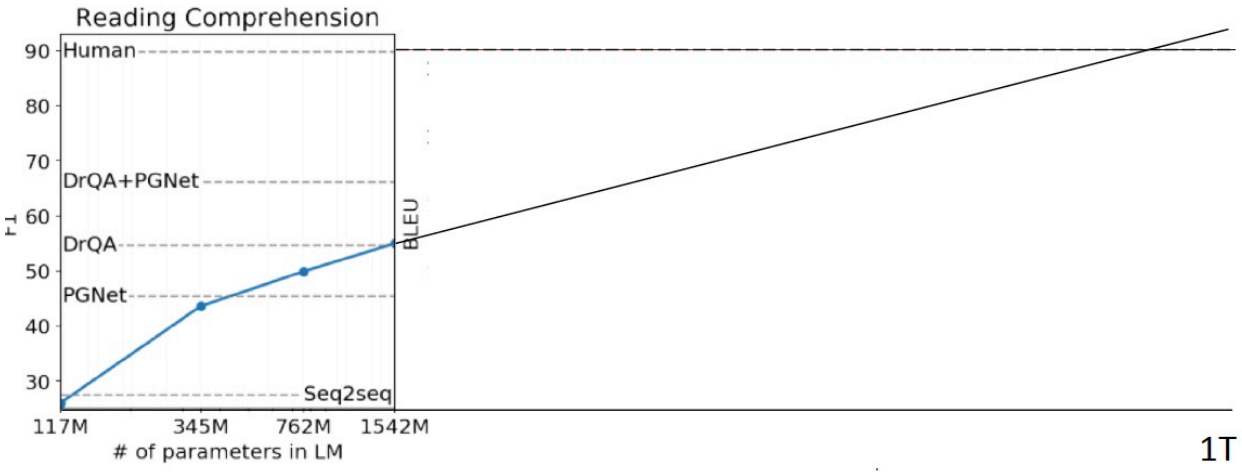
- 由于数据集中存在一些翻译的例子
  - 法语习语及其翻译
  - 法语引用及其翻译

GPT-2 Question Answering

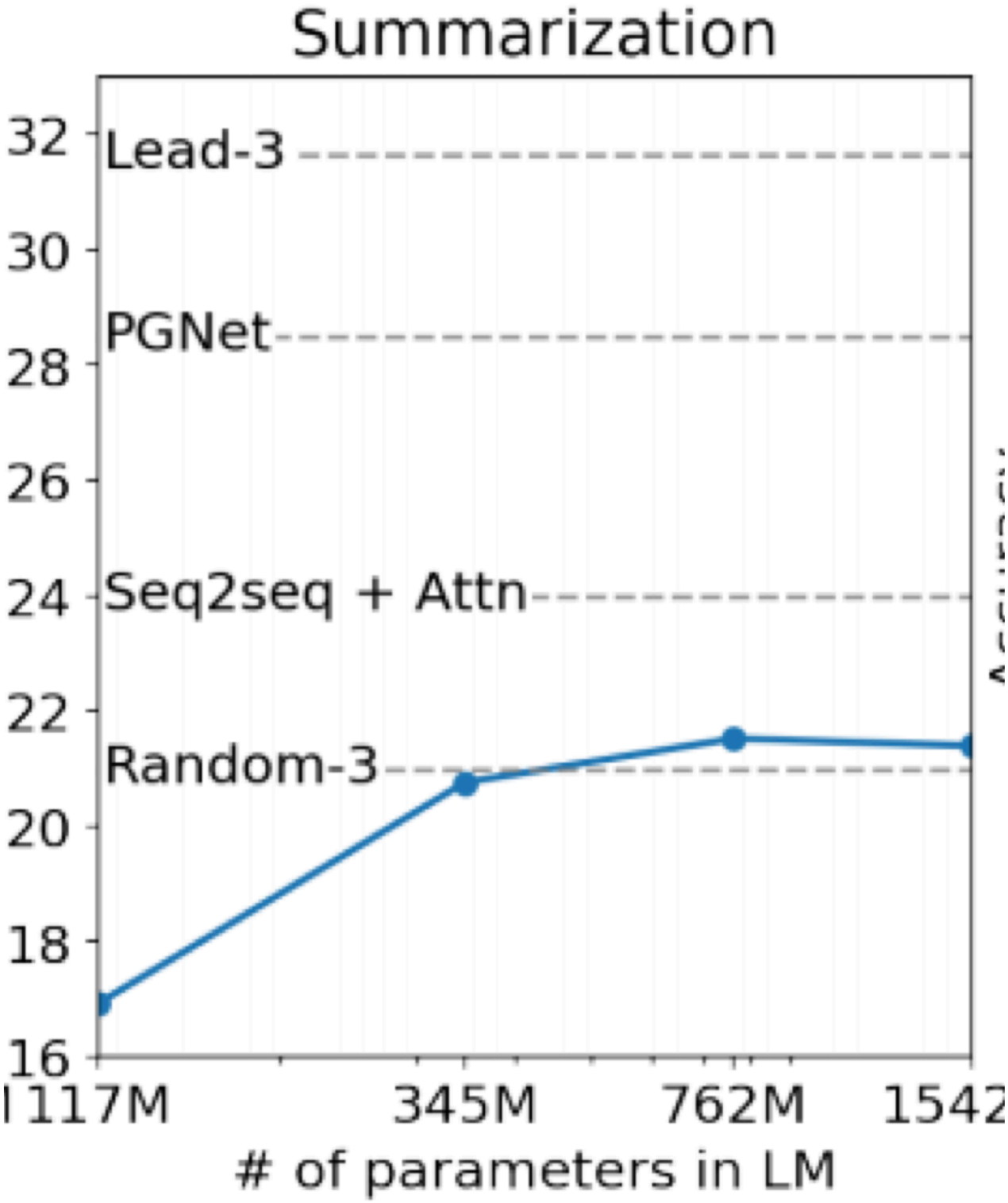
- Simple baseline: 1% accuracy
- GPT-2: ~4% accuracy
- Cherry-picked most confident results 精选出最自信的结果

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%

What happens as models get even bigger?



- 对于一些任务，性能似乎随着 log(模型大小) 的增加而增加
- 但如下图所示趋势并不明朗



GPT-2 Reaction

Some arguments for release:

- This model isn't much different from existing work
- Not long until these models are easy to train
  - And we're already at this point for images/speech
- Photoshop
- Researchers should study this model to learn defenses
- Dangerous PR Hype
- Reproducibility is crucial for science
- …

Some arguments against:

- Danger of fake reviews, news comments, etc.
  - Already done by companies and governments
- Precedent
  - Event if this model isn't dangerous, later ones will be even better
- Smaller model is being released
- ….

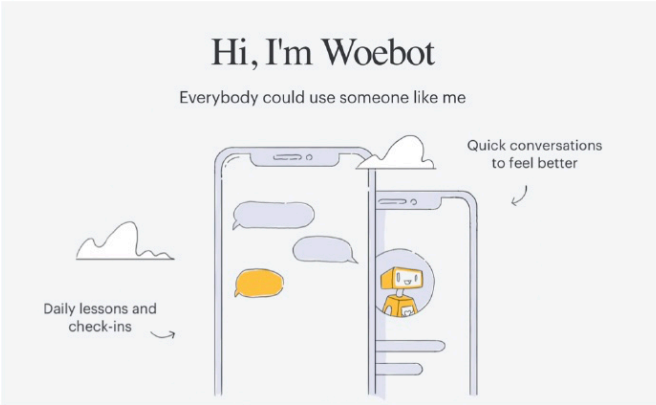
- NLP专家应该做这些决定吗?
  - 计算机安全专家?
  - 技术和社会专家?
  - 道德专家?
- 需要更多的跨学科科学
- 许多NLP具有较大社会影响的例子，尤其是对于偏见/公平

High-Impact Decisions

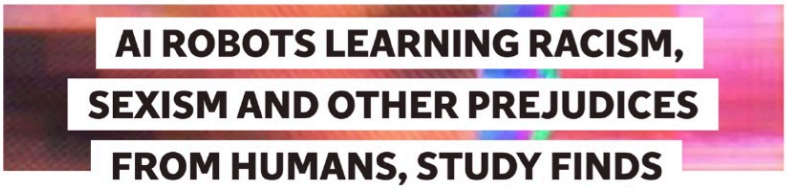
- 越来越感兴趣用NLP帮助高影响力的决策
  - 司法判决
  - 招聘
  - 等级测试
- 一方面，可以快速评估机器学习系统某些偏见
- 然而，机器学习反映了训练数据
  - 甚至放大偏见…这可能导致更偏向数据的创建

Chatbots

- Potential for positive impact

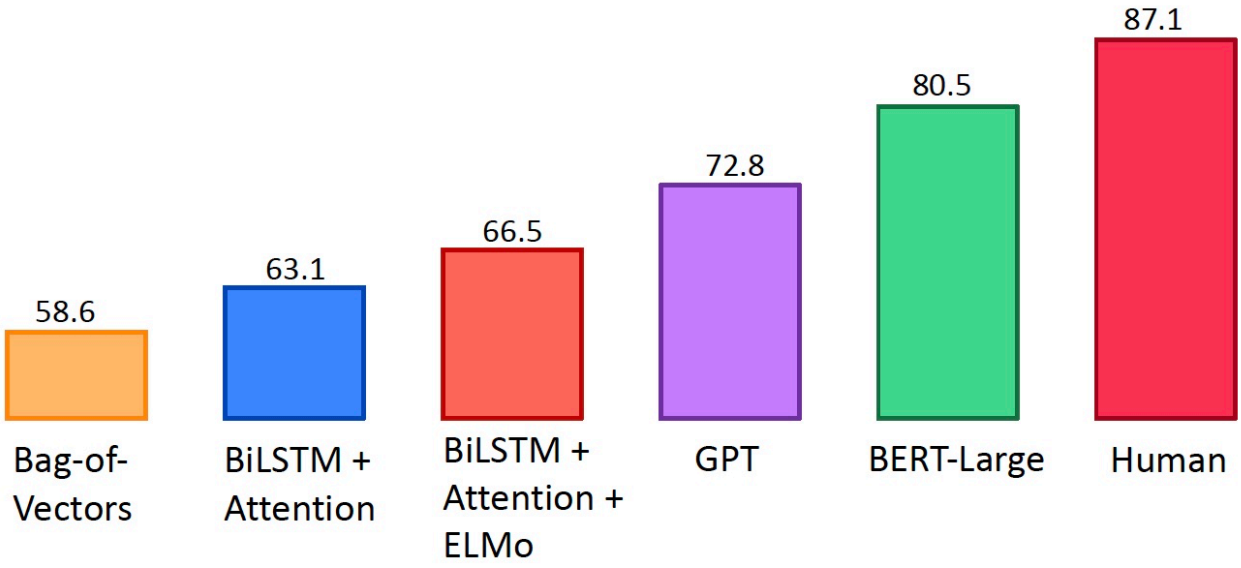


- But big risks



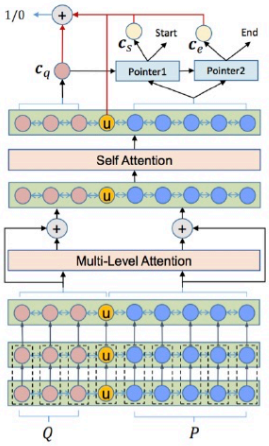
What did BERT "solve" and what do we work on next?

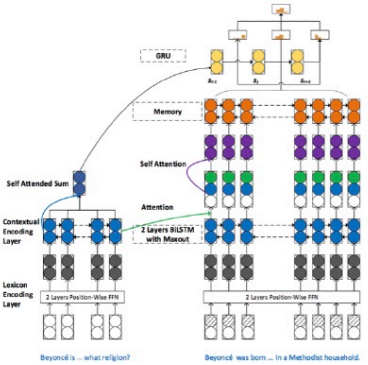
GLUE Benchmark Results

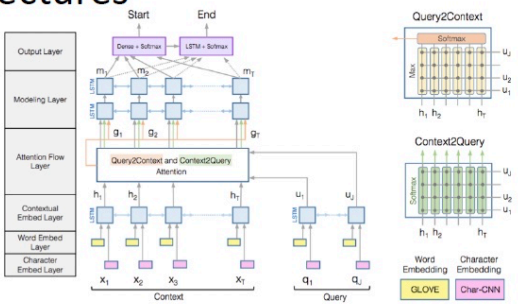


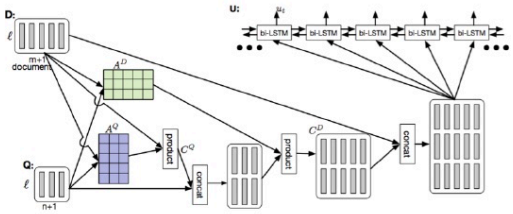
The Death of Architecture Engineering?

Some SQuAD NN Architectures

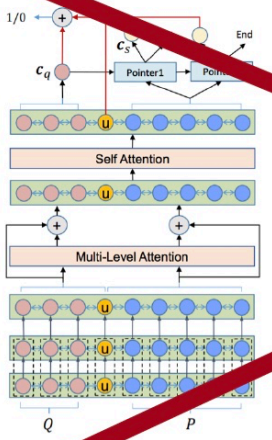


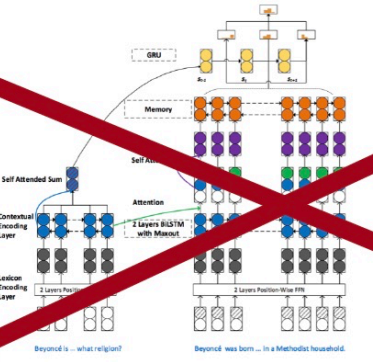


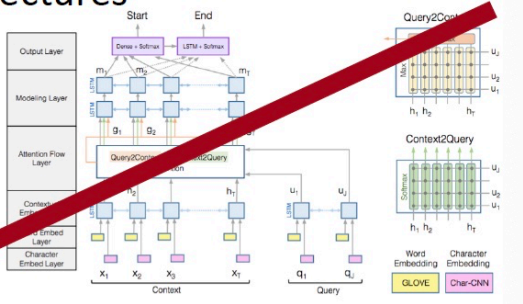


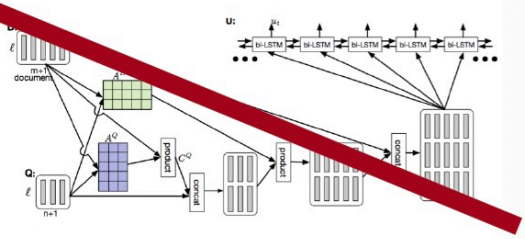


Some SQuAD NN Architectures









Attention Is All You Need

	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	**86.673**	**89.147**
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
3 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
4 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
5 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
5 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
6 Mar 04, 2019	SemBERT (ensemble model) *Shanghai Jiao Tong University*	83.243	85.821
6 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892
7 Jan 10, 2019	BERT + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	82.972	85.810

- 花费六个月来研究 体系结构的设计，得到了 1 F1 的提升
- 或知识让 BERT 变得 3 倍大小，得到了 5 F1 的提升
- SQuAD 的 TOP20 参赛者都是用了 BERT

Harder Natural Language Understanding

- 阅读理解
  - 在长文档或多个文档
  - 需要多跳推理
  - 在对话中定位问答
- 许多现有阅读理解数据集的关键问题：人们写问题时看着上下文
  - 不现实的
  - 鼓励简单的问题

QuAC: Question Answering in Context

Section:  Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ↪ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

> Choi et al., 2018

- 学生问问题，老师回答的对话

    - 教师看到维基百科文章主题，学生不喜欢
- 仍然和人类水平有很大差距

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
 March 7, 2019	BERT w/ 2-context (single model) *NTT Media Intelligence Labs*	**64.9**	60.2	**6.1**
2 Feb 21, 2019	GraphFlow (single model) *Anonymous*	**64.9**	**60.3**	5.1
3 Sep 26, 2018	FlowQA (single model) *Allen Institute of AI* https://arxiv.org/abs/1810.06683	64.1	59.6	5.8
4 Jan 30, 2019	BERT + History Answer Embedding (single model) *Anonymous*	62.4	57.8	5.1
5 Aug 20, 2018	BiDAF++ w/ 2-Context (single model) *baseline*	60.1	54.8	4.0
6 Aug 20, 2018	BiDAF++ (single model) *baseline*	50.2	43.3	2.2

HotPotQA

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HotpotQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

> Zang et al., 2018

- 设计要求多跳推理
- 问题在多个文档
- Human performance is above 90 F1

	Model	Code	Ans	
			EM	F_1
1 Nov 21, 2018	QFE (single model) *NTT Media Intelligence Laboratories*		53.86	68.06
2 Mar 4, 2019	GRN (single model) *Anonymous*		52.92	66.71
3 Mar 1, 2019	DFGN + BERT (single model) *Anonymous*		55.17	68.49
4 Mar 4, 2019	BERT Plus (single model) *CIS Lab*		**55.84**	**69.76**
5 Oct 10, 2018	Baseline Model (single model) *Carnegie Mellon University, Stanford University, & Universite de Montreal* (Yang, Qi, Zhang, et al. 2018)		45.60	59.02
- Feb 27, 2019	DecompRC (single model) *Anonymous*		55.20	69.63

Multi-Task Learning

BERT + Multi-task

Rank	Name	Model	URL	Score
1	GLUE Human Baselines	GLUE Human Baselines		87.1
2	王玮	ALICE large (Alibaba DAMO NLP)		83.0
3	Microsoft D365 AI & MSR AI	MT-DNNv2 (BigBird)		83.0
4	Jason Phang	BERT on STILTs		82.0
5	Jacob Devlin	BERT: 24-layers, 16-heads, 1024-hid		80.5

- NLP的另一个前沿是让一个模型执行许多任务。GLUE 和 DecaNLP是最近的例子
- 在BERT的基础上，多任务学习产生了改进

Low-Resource Settings

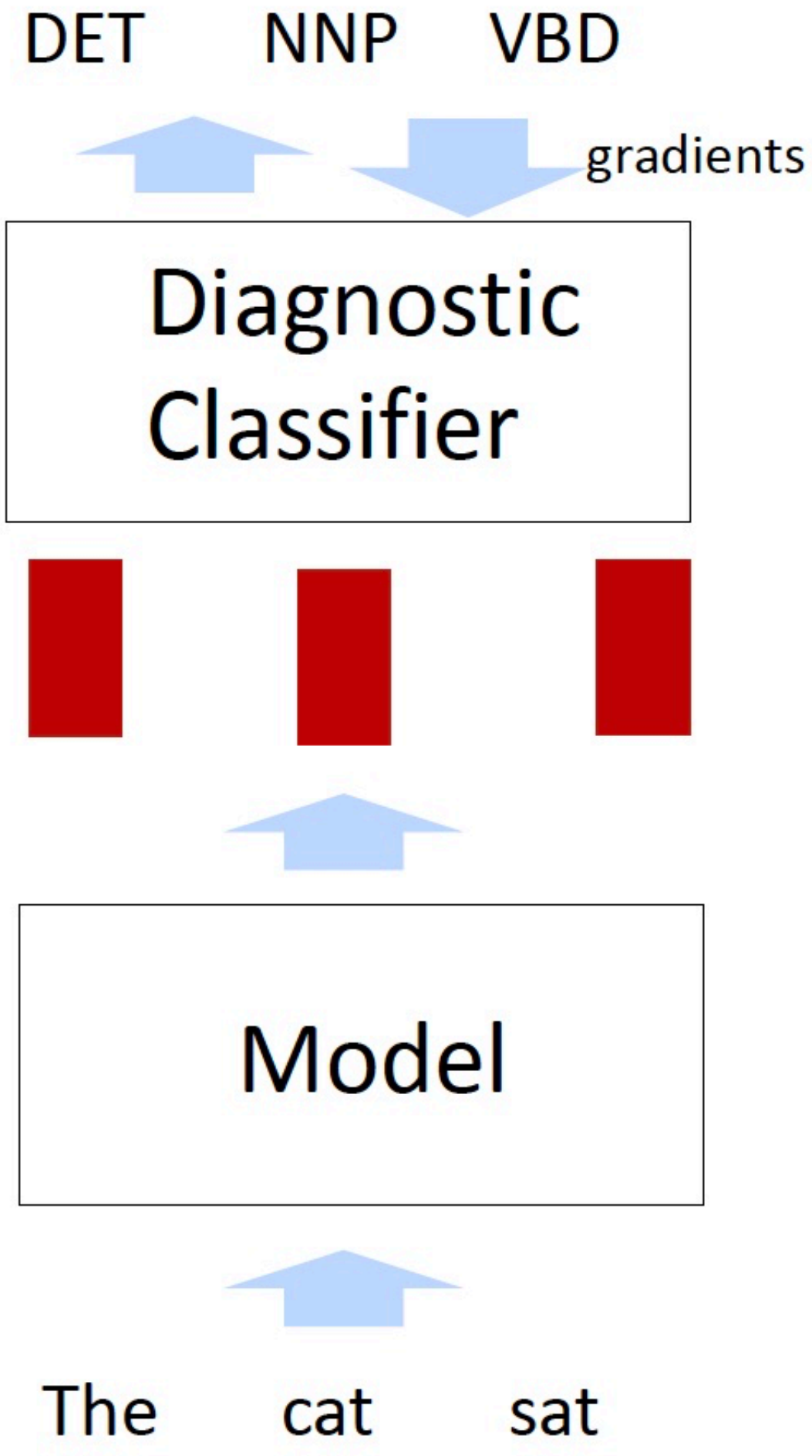
- 不需要很多计算能力的模型(不能使用BERT)
  - 为移动设备尤其重要

- 低资源语言
- 低数据环境(few shot learning 小样本学习)
    - ML 中的元学习越来越受欢迎

Interpreting/Understanding Models

- 我们能得到模型预测的解释吗?
- 我们能理解模型，例如BERT知道什么和他们为什么工作这么好?
- NLP中快速增长的地区
- 对于某些应用程序非常重要(如医疗保健)

Diagnostic/Probing Classifiers

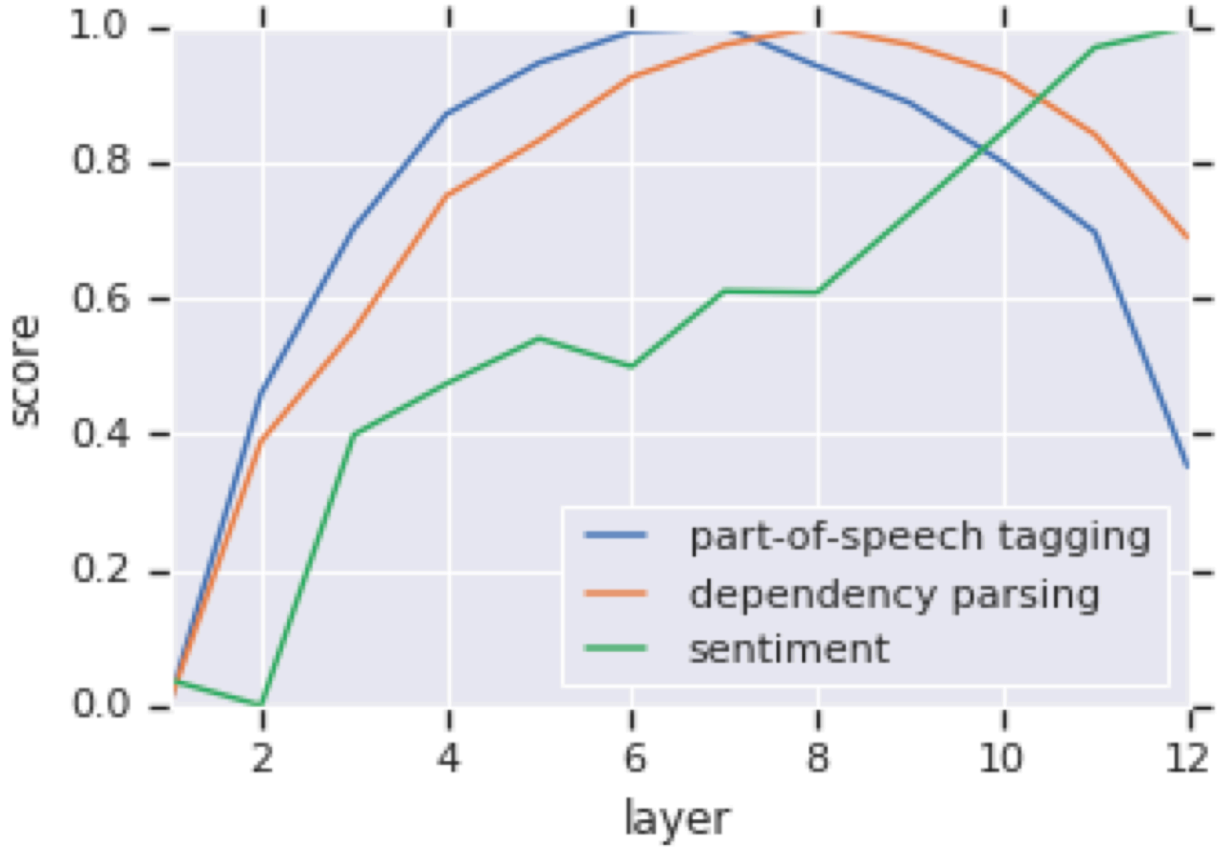


- 看看模型知道什么语言的信息

- 诊断分类器需要表示一个模型(例如BERT)作为输入，并做一些任务
- 只有诊断分类器被训练
- 诊断分类器通常非常简单(例如，单个softmax)。否则他们不通过模型表示来自省会学会完成任务
- 一些诊断任务

POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. $\rightarrow$ NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. $\rightarrow$ VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ $\rightarrow$ nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. $\rightarrow$ Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. $\rightarrow$ Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy... $\rightarrow$ {awareness, existed_after, ...}
Coref.$^{\text{O}}$	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. $\rightarrow$ True
Coref.$^{\text{W}}$	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. $\rightarrow$ True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. $\rightarrow$ False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. $\rightarrow$ Cause-Effect(e_2, e_1)

Diagnostic/ Probing Classifiers: Results



- 低层的 BERT 在低层的任务中表现更好

NLP in Industry

- Dialogue
  - Chatbots
  - Customer service

- Healthcare
  - Understanding health records
  - Understanding biomedical literature

- NLP是快速增长的行业。尤其是两大领域：
- 对话
  - 聊天机器人
  - 客户服务
- 健康
  - 理解健康记录
  - 理解生物医学文献

Conclusion

- 在过去的5年里，由于深度学习，进步很快
- 由于较大的模型和更好地使用无标记数据，在去年有了更大的进展
  - 是在NLP领域的激动人心的时刻
- NLP是正逐渐对社会产生巨大影响力，使偏差和安全等问题越来越重要

Reference

以下是学习本课程时的可用参考书籍：

《基于深度学习的自然语言处理》（车万翔老师等翻译）

《神经网络与深度学习》

以下是整理笔记的过程中参考的博客：

斯坦福CS224N深度学习自然语言处理2019冬学习笔记目录 (课件核心内容的提炼，并包含作者的见解与建议)

斯坦福大学 CS224n自然语言处理与深度学习笔记汇总 {>>这是针对note部分的翻译<<}