- 第17课：多任务学习 （举例decaNLP）（客座讲座）
- 第18课：成分句法分析（Constituency Parsing）、树递归神经网络（Tree Recursive Neural Networks）
- 第19课：视觉和语言AI中的偏见（Bias）（客座讲座）
- 第20课：NLP深度学习的未来

第17课：多任务学习 （举例decaNLP）（客座讲座）

NLP和AI的发展经过

- 基于特征工程的机器学习
- 基于特征学习的深度学习
- 为单一任务的深度结构工程
- 未来是什么？单一的多任务模型？

单一任务学习在过去几年中发展很好，在数据集足够大情况下可以达到局部最优，但一般都是从随机开始训练，或者仅有部分预训练。

预训练和共享知识

- 计算机视觉领域，ImageNet+CNN获得很大成功，图片分类是机器视觉领域的阻碍级别任务
- NLP领域开始有了一些预训练Word2Vec、GloVe、CoVe、ELMo、BERT，没有单一的阻碍级别任务
- 然而NLP领域没有那么多的权重和模型共享，因为：
- NLP有多种类型的推理
- NLP需要短期记忆和长期记忆
- NLP领域被划分为很多独立的任务
- 语言自然具有监督性，无法找到单一的无监督任务解决所有问题

统一的NLP多任务模型

- 多任务学习就是NLP领域阻碍级别的任务
- 多任务学习可以用于传播知识
- 更容易解决新问题

NLP任务汇总到统一框架内

- 序列标注：NER、Aspect Specific Sentiment
- 文本分类：对话状态跟踪、情感分类
- Seq2seq：机器翻译、摘要、问答

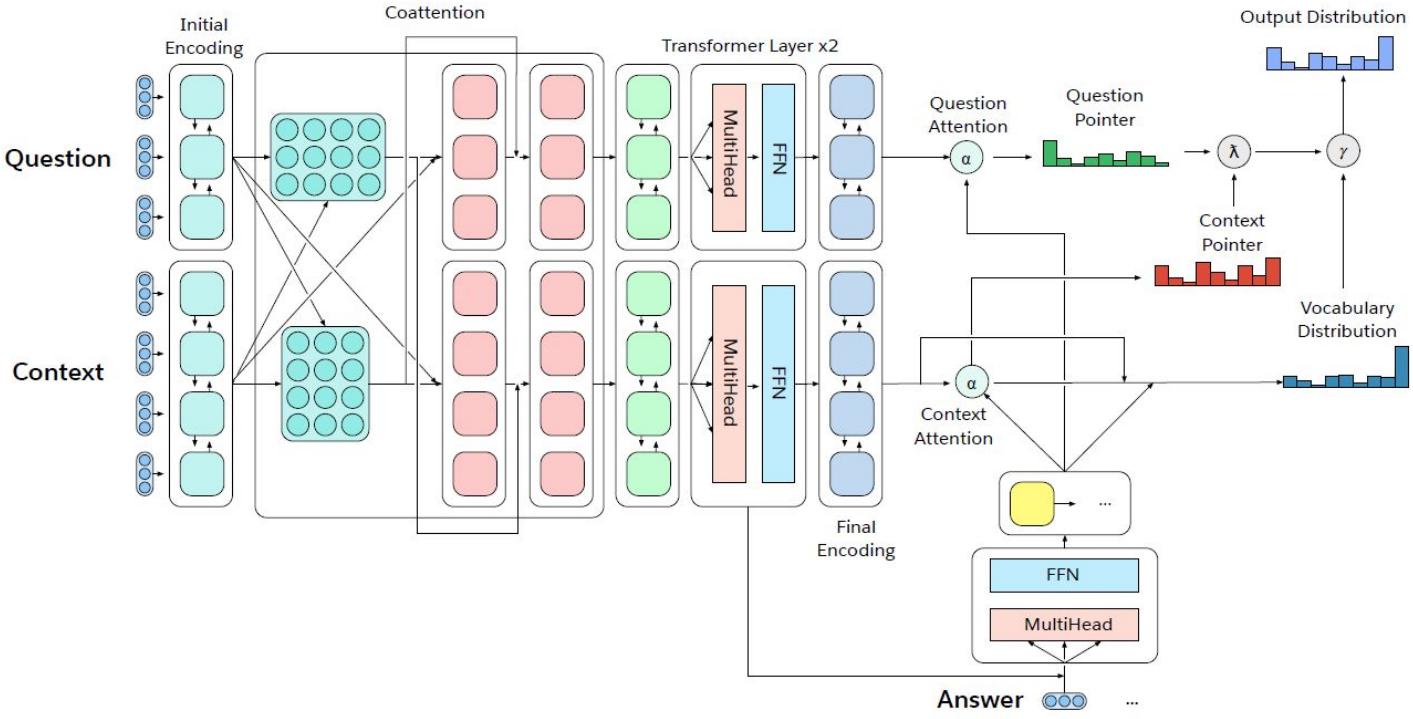
NLP的三个超级任务：语言模型LM、问答QA、对话Dialogue

NLP十项全能：decaNLP

- 问答、机器翻译、摘要、自然语言推理、情感分类、语义角色标签、关系提取、对话、语义分析、常识推理
- 元-监督学习：{x, y} 到 {x, t, y}

decaNLP中提出的一种多任务问答网络

- initial encoding：固定的GloVe和字符n-gram嵌入 -> 线性连接 -> 共享BiLSTM，带skip connection
- Coattention：一个序列对另一个序列进行attention求和，加上skip connection
- Transformer Layer x2：两个单独的BiLSTM用于降维，两个Transformer层，再接上BiLSTM
- Answer：一个自回归decoder，包括固定的GloVe和字符n-gram嵌入，两个Transformer层和一个LSTM层，attention注意到encoder最后三个层的输出
- Question Attention, Context Attention, Question Pointer, Context Pointer, Vocabulary Distribution： LSTM decoder状态，用于计算context和query的注意力分布，用于形成Pointer
- Output Distribution：gamma决定是否从外部词汇表挑选，lambda决定是否从context或query中决定



decaNLP架构

decaNLP训练任务

- 采用诸多任务轮流训练
- 反课程顺序预训练：即按照难度逆序训练

Evaluation	Dataset	Metric
Question Answering	SQuAD	nF1
Machine Translation	IWSLT En — De	BLEU
Summarization	CNN/DailyMail	ROUGE
Natural Language Inference	MultiNLI	EM
Sentiment Analysis	SST2	EM
Semantic Role Labeling	QA-SRL	nF1
Relation Extraction	QA-ZRE	cF1
Goal-Oriented Dialogue	WOZ	dsEM
Semantic Parsing	WikiSQL	lfEM
Pronoun Resolution	Winograd Schemas	EM

decaNLP训练任务

参考文献

- Multitask Learning
  - Collobert and J. Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In ICML, 2008.
  - M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. S. Corrado, M. Hughes, and J. Dean. Google’ s multilingual
  - neural machine translation system: Enabling zero-shot translation. TACL, 5:339–351, 2017.
  - M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser. Multi-task sequence to sequence learning. CoRR, abs/1511.06114, 2015a.
  - L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit. One model to learn them all. CoRR, abs/1706.05137, 2017.
- Model
  - A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In ACL, 2017.
- Training
  - Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In ICML, 2009.

第18课：成分句法分析（Constituency Parsing）、树递归神经网络（Tree Recursive Neural Networks）

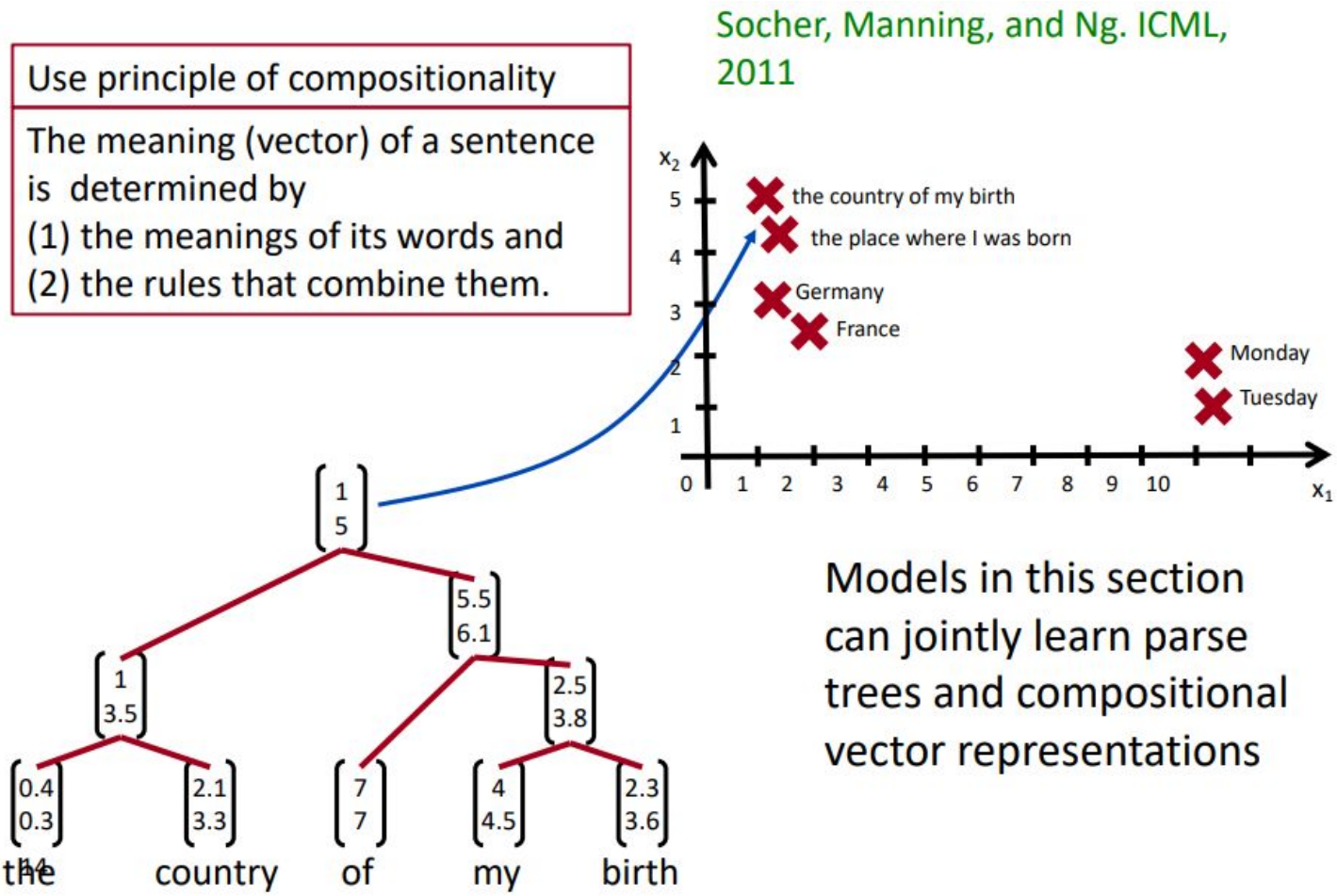
> 由于循环神经网络和递归神经网络表面上都可以简写成RNN，所以要注意上下文，在本节内容中基本都直接称呼为RNN，但在更大讨论环境下通常携程RvNN。

语言并不只是一个单词的列表，而是语义的合成。

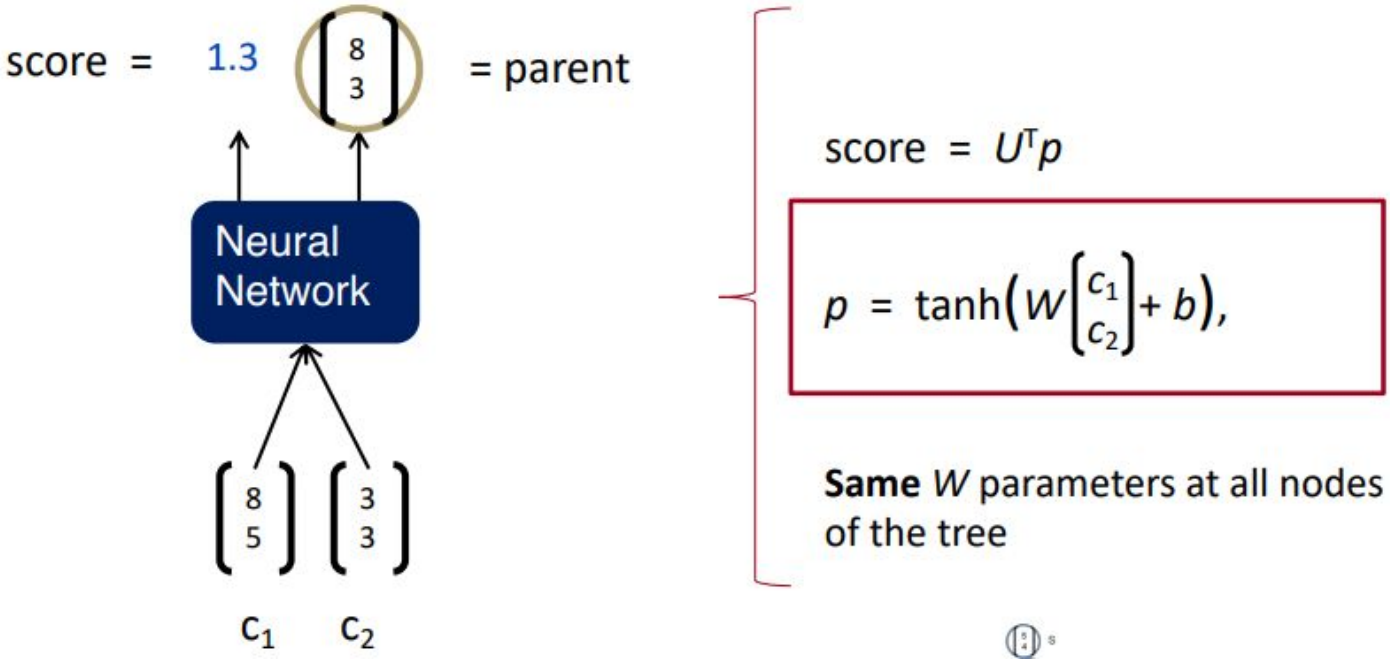
语言是递归的，虽然有一些认知争议，但递归确实是一种自然的语言描述方式

第一版：简单的TreeRNN

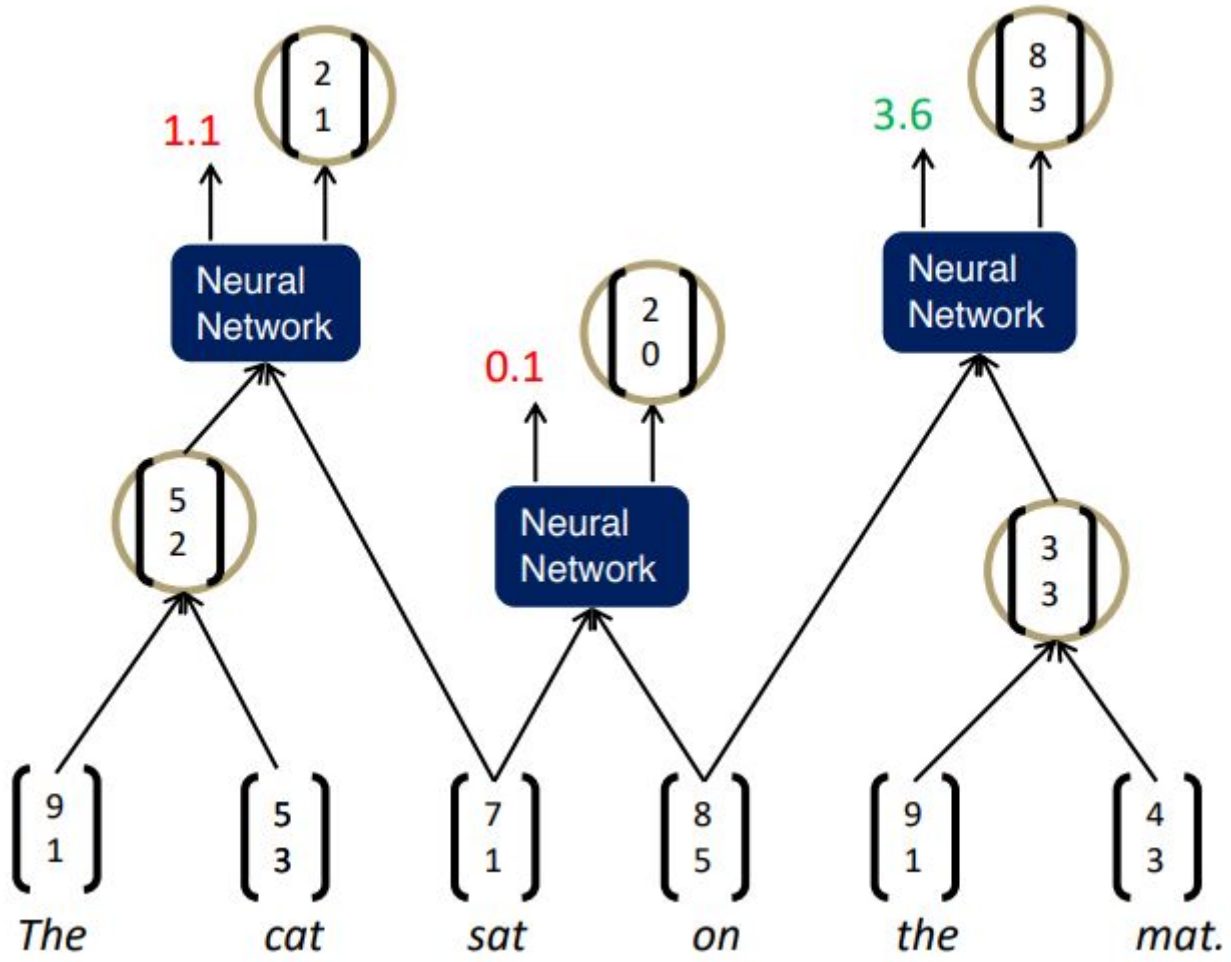
针对长句子，同样将其映射到一个相同的向量空间中：



通过两个子节点表示[c1,c2]，计算父节点表示p，以及可信分数score：



采用对score的贪婪选择，形成整个句法分析树：



Max-Margin框架：损失函数。

- $s(x, y)$ 表示句子x在解析树y情况下，所有节点score的加和
- $\Delta(y, y_i)$ Δ(y, yi)惩罚所有错误决定

$$J = \sum_i s(x_i, y_i) - \max_{y \in A(x_i)} \left(s(x_i, y) + \Delta(y, y_i) \right)$$

Max-Margin

自然语言句法分析和图像分解表达，都可以用递归神经网络解决：

- Parsing Natural Scenes and Natural Language with Recursive Neural Networks (Socher et al. ICML 2011)

RvNN的反向传播

- 累加所有节点的导数
- 向子节点切分导数
- 每个节点用自己score差值和父节点传播来的差值进行加和，再向后传播

$$\delta^{(l)} = \left((W^{(l)})^T \delta^{(l+1)} \right) \circ f'(z^{(l)}),$$

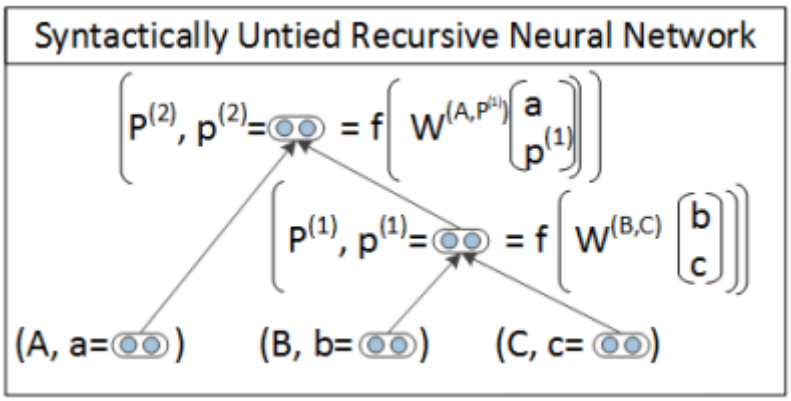
$$\frac{\partial}{\partial W^{(l)}} E_R = \delta^{(l+1)} (a^{(l)})^T + \lambda W^{(l)}$$

简单的TreeRNN能够捕捉一些句法现象，但无法处理更复杂的句法组合或者处理更长的句子。

第二版：Syntactically-Untied RNN [Socher, Bauer, Manning, Ng 2013]

采用上下文无关文法（CFG），针对子节点尝试融合时的矩阵采用不同的矩阵。

为了提升速度，修改采用PCFG来快速计算确定一个候选子集。



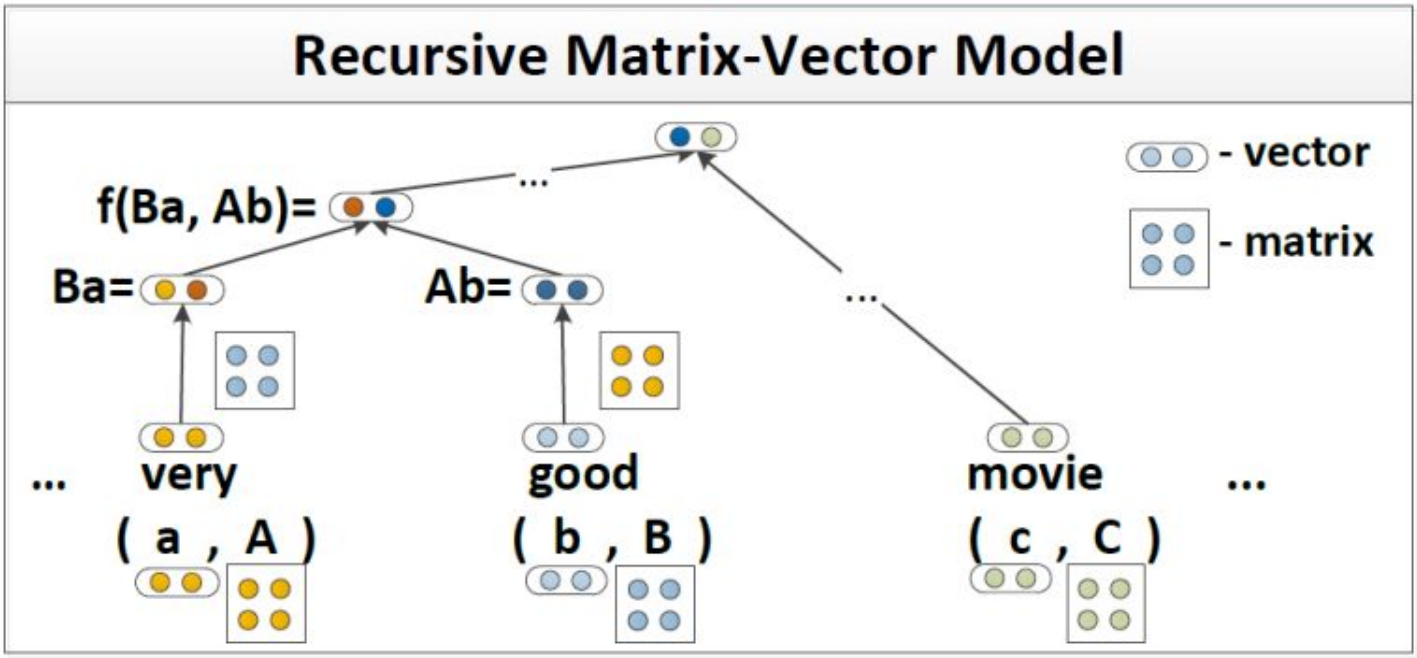
SU-RNN / CVG [Socher, Bauer, Manning, Ng 2013]

第三版：Compositionality Through Recursive Matrix-Vector Spaces [Socher, Huval, Bhat, Manning, & Ng, 2012]

采用更复杂的融合算法：

$$p = \tanh\left(W\begin{bmatrix}c_1\\c_2\end{bmatrix}+ b\right)$$

$$p = \tanh\left(W\begin{bmatrix}C_2c_1\\C_1c_2\end{bmatrix} + b\right)$$



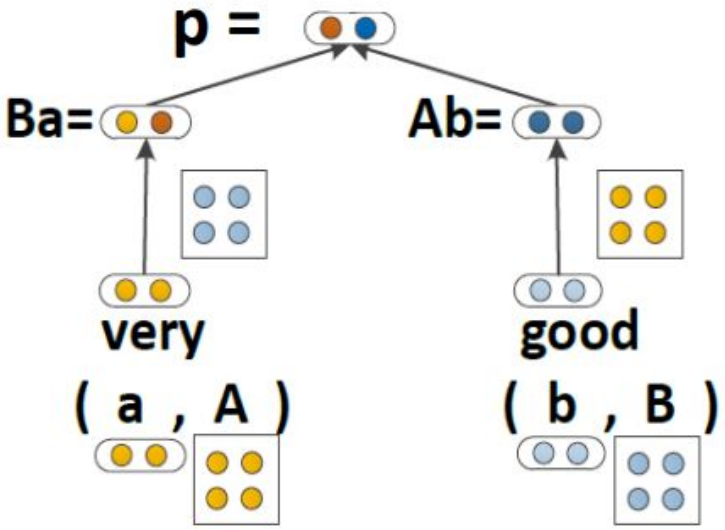
Matrix-vector RNNs [Socher, Huval, Bhat, Manning, & Ng, 2012]

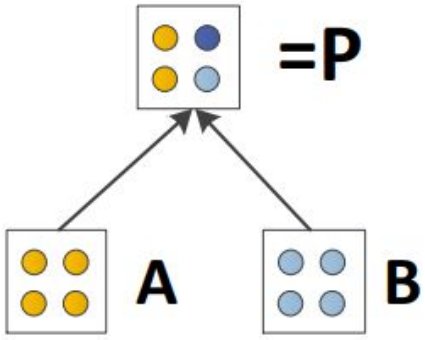
很好地处理长句子中的两个词之间的句法关系分类。

$$p = f\left(W\begin{bmatrix} Ba \\ Ab \end{bmatrix}\right)$$

$$P = g(A,B) = W_M\begin{bmatrix} A \\ B \end{bmatrix}$$

$$W_M \in \mathbb{R}^{n\times 2n}$$

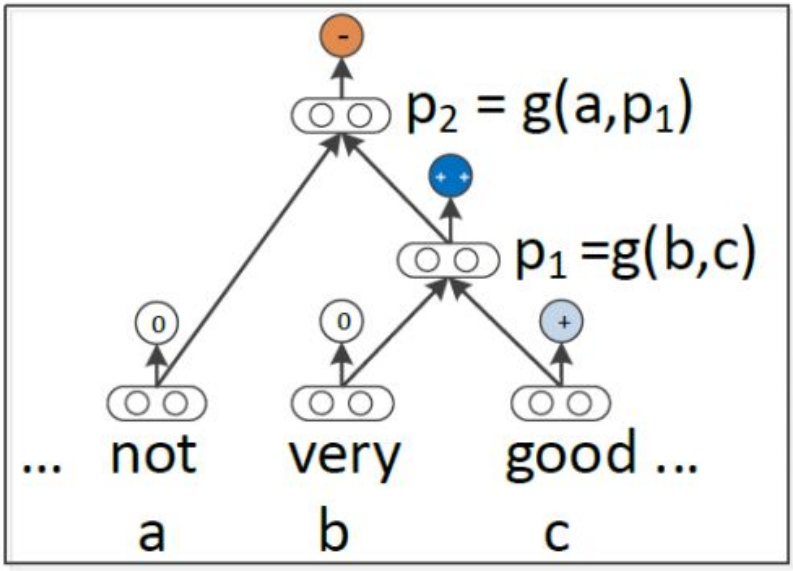


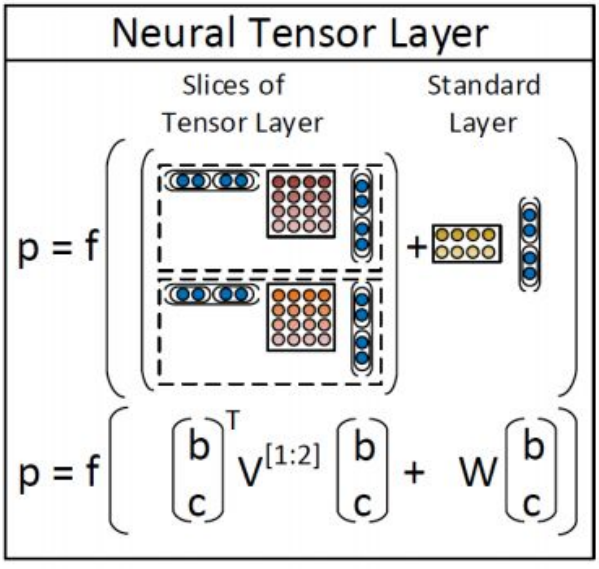


第四版：Recursive Neural Tensor Network [Socher, Perelygin, Wu, Chuang, Manning, Ng, and Potts 2013]

允许两个子节点的表达向量用乘法相互影响

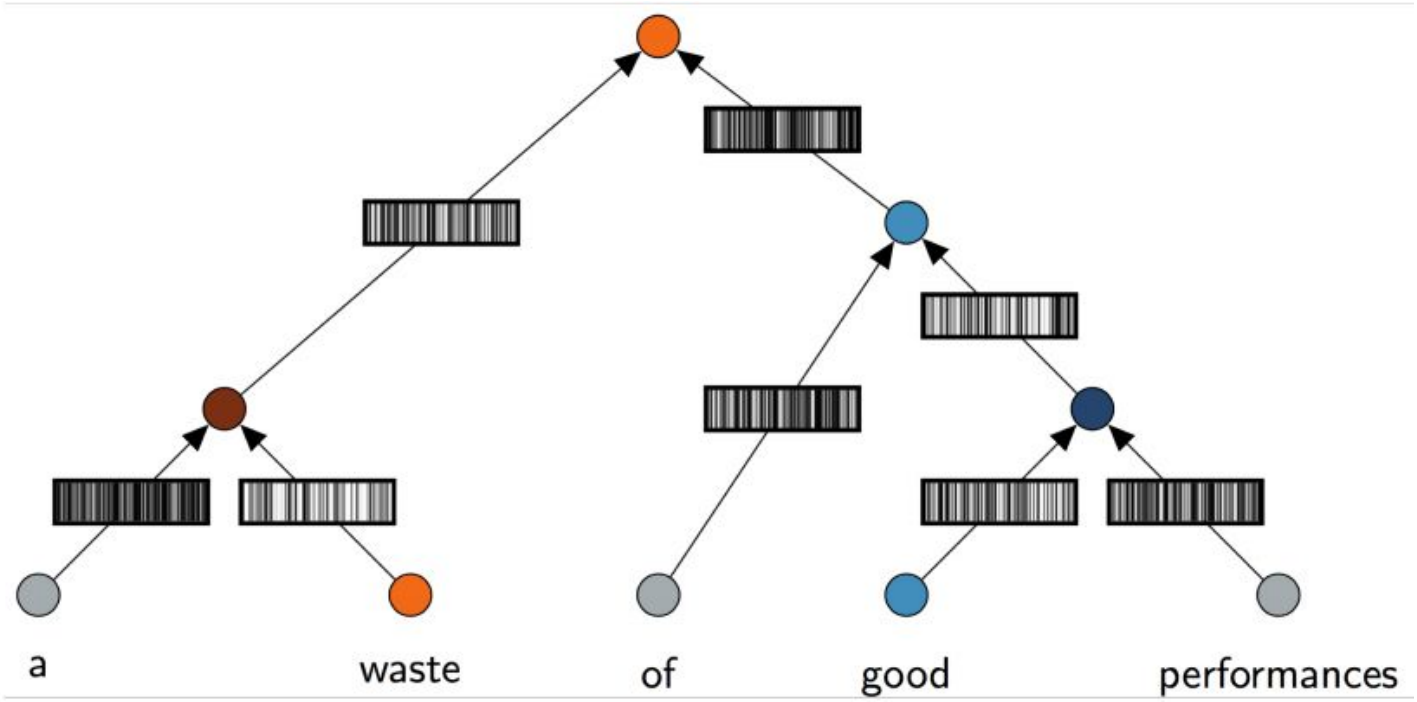
在情感分类方面取得提升





第五版：Improving Deep Learning Semantic Representations using a TreeLSTM [Tai et al., ACL 2015; also Zhu et al. ICML 2015]

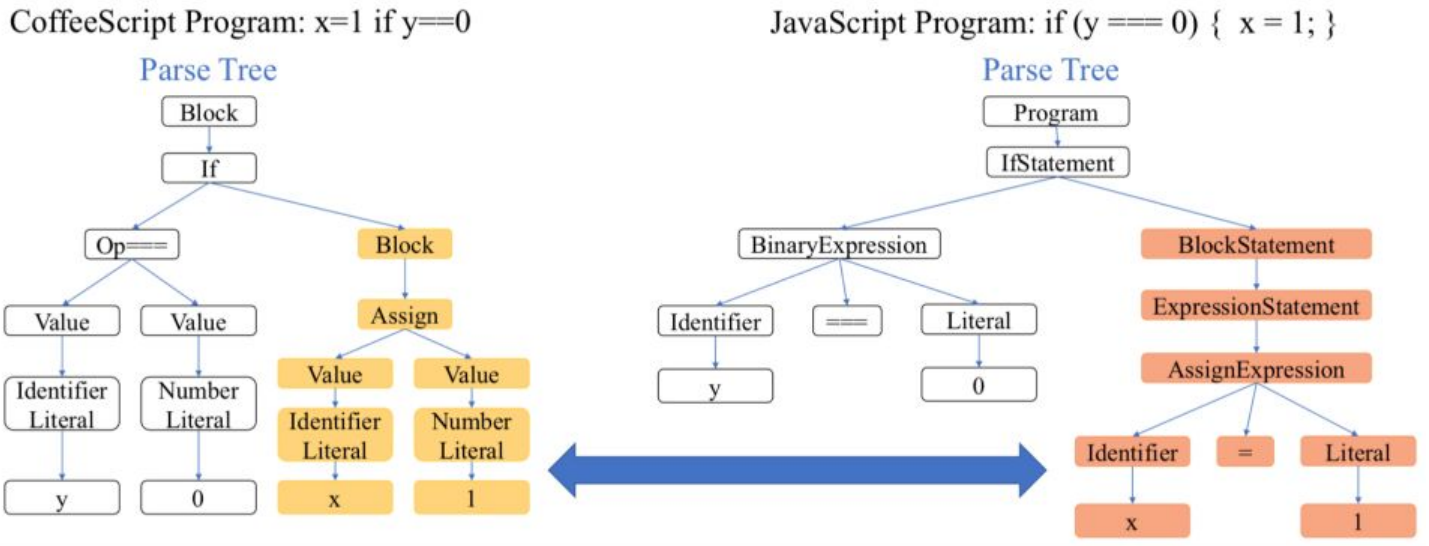
利用LSTM中的forget门，选择哪些信息要忘掉，哪些要保留。



（图中深色竖条表示遗忘，更多白色表示保留）

QCD-Aware Recursive Neural Networks for Jet Physics Gilles [Louppe, Kyunghun Cho, Cyril Becot, Kyle Cranmer (2017)]

Tree-to-tree Neural Networks for Program Translation [Chen, Liu, and Song NeurIPS 2018]

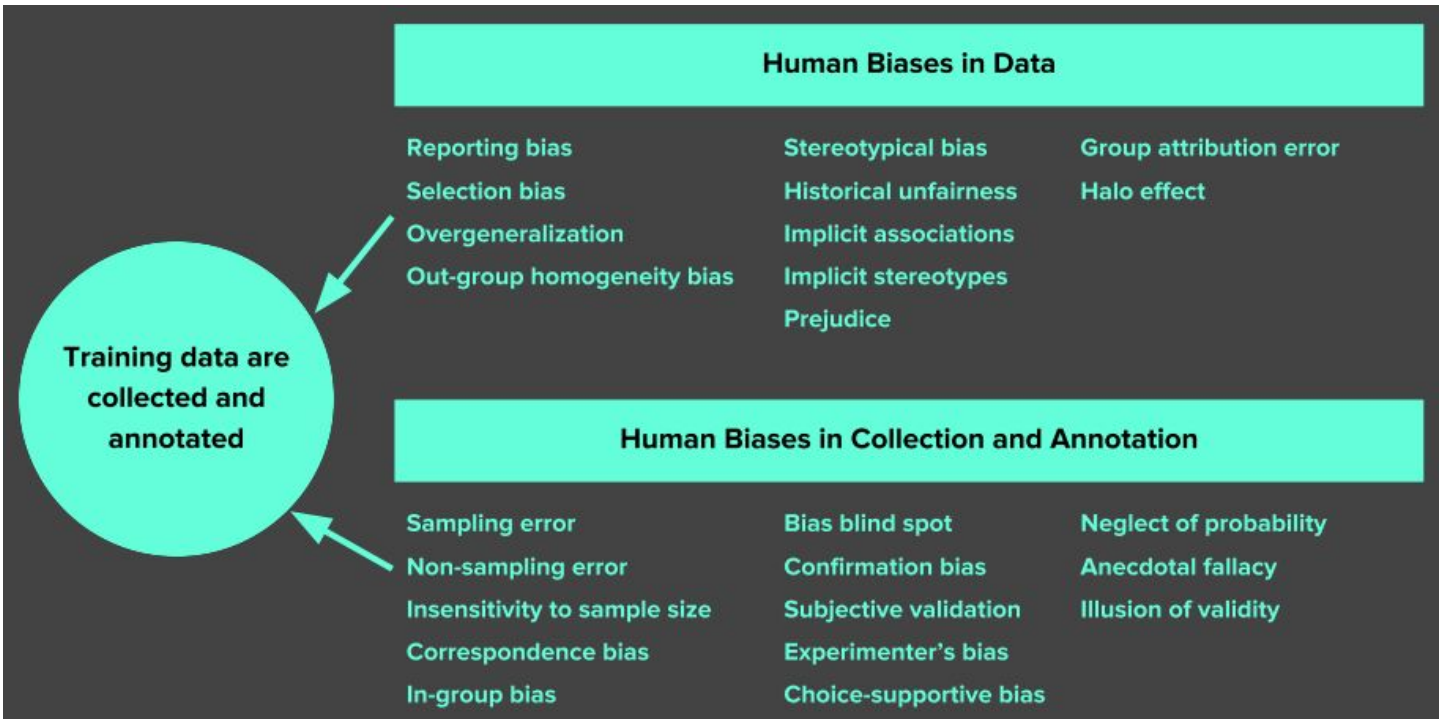


第19课：视觉和语言AI中的偏见（Bias）（客座讲座）

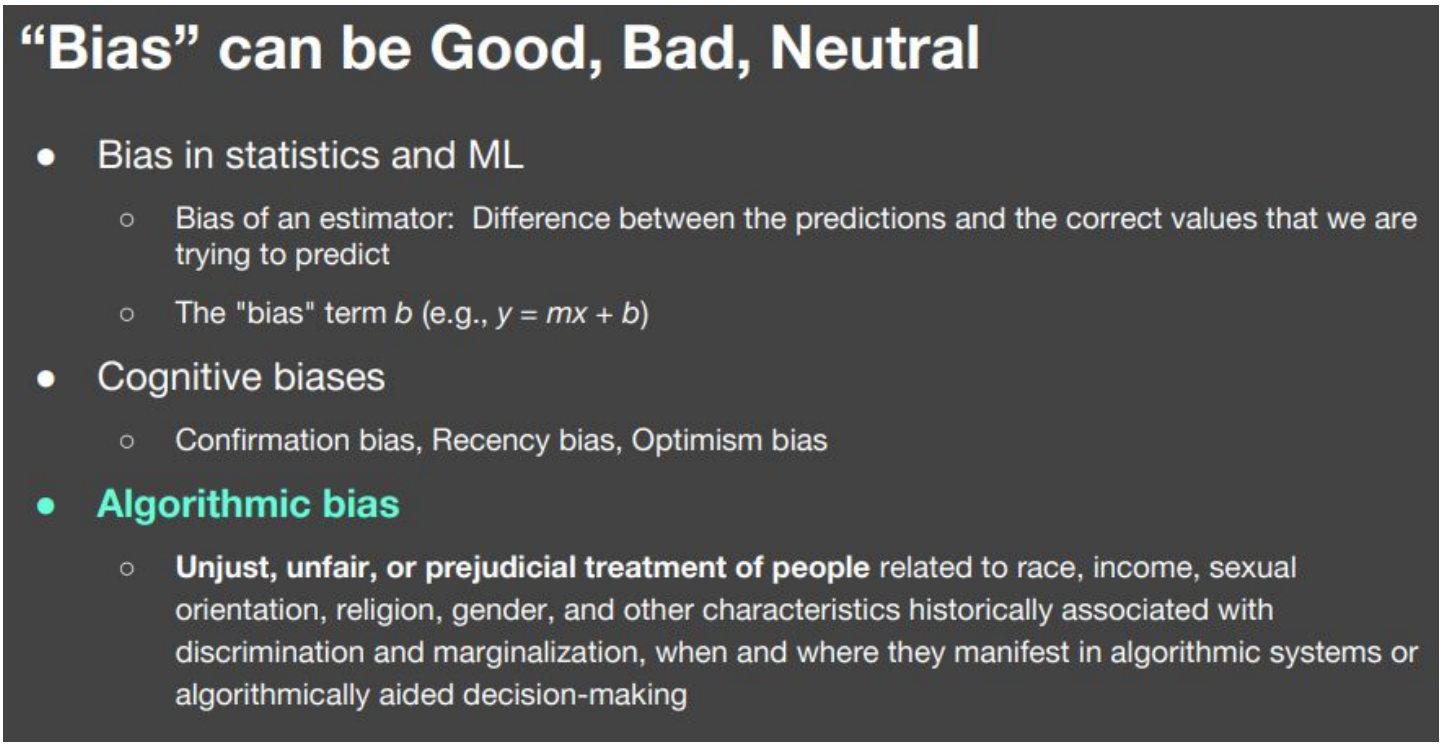
人看到香蕉，能分析出很多结论，但很难分析出“黄色的香蕉”，因为黄色香蕉是最典型的香蕉，在认知中会当成默认情况。

对于“doctor”，无论男女还是女权主义者都会首先想到是男的，而如果是女医生反而要说“female doctor”。

准备数据时的各类偏差，并且偏差会在整个机器学习系统中存在、迭代、循环：



产生的各类偏差结果：



近期公开的针对犯罪率预测的系统、针对判断LGBT的系统，可以发现很多算法偏差：

- 过度泛化
- 反馈循环

- 相关性谬误

如何评估自动偏差?

- 通过划分不同子群体，分别计算每个群体之中的TP/FP/FN/TN和相关比例，看是否相同
- 重点评判FP和FN的可接受情况

解决方法

- 尽量收集更加全面、公平的数据
- 运用机器学习方式减少偏差
- 减少刻板印象、性别歧视、种族歧视等有问题的输出
- 包容：增加一些渴望变量的信号
- 通过多任务学习增加包容性
- 多任务对抗学习

[**Measuring and Mitigating Unintended Bias in Text Classification**, AIES, 2018 and FAT*, 2019]

- 发现原数据对LGBT有攻击性，补充了很多无攻击性的数据
- 研究有攻击性和无攻击性数据的ROC-AUC，发现和测量偏差

第20课：NLP深度学习的未来

五年之前，没有Seq2seq，没有attention，没有大规模问答/阅读理解数据集，也没有TensorFlow或PyTorch。

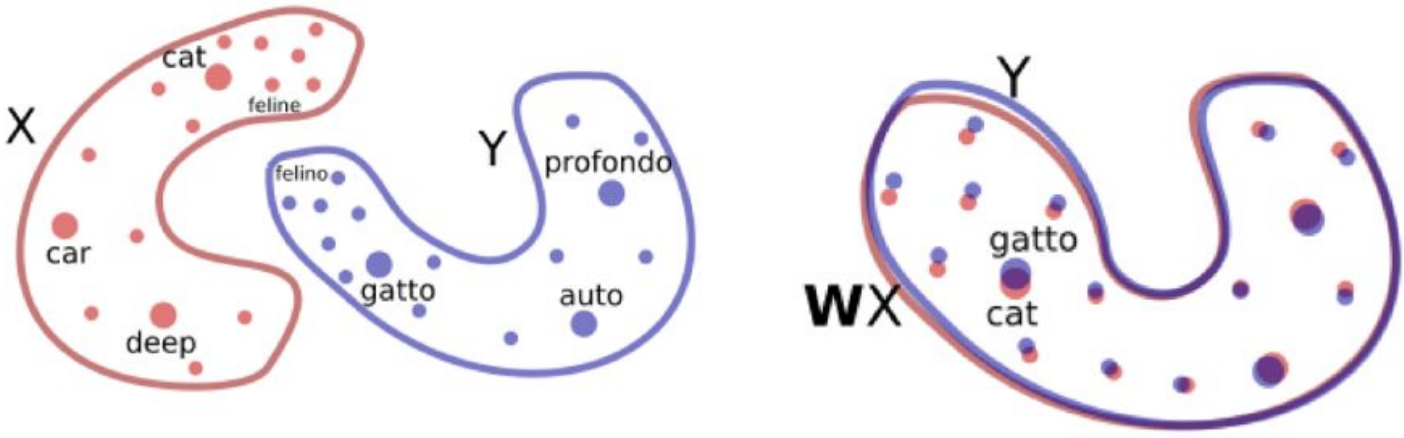
使用未标注数据进行机器翻译

- 预训练：比如分别预训练两个语言的LM，然后拼接成encoder-decoder，采用双语翻译语料训练
- 自训练：对无标签数据用现有模型进行标签计算，获得嘈杂的训练数据
- 反向翻译：通过正向翻译的结果，对反向翻译进行训练
- 无监督词汇翻译：在没有双语翻译数据的情况下，分别对两种语言内部进行word embedding，然后寻找对应关系

无监督词汇翻译：

- 对两种语言内部进行word embedding后，需要进行转换匹配

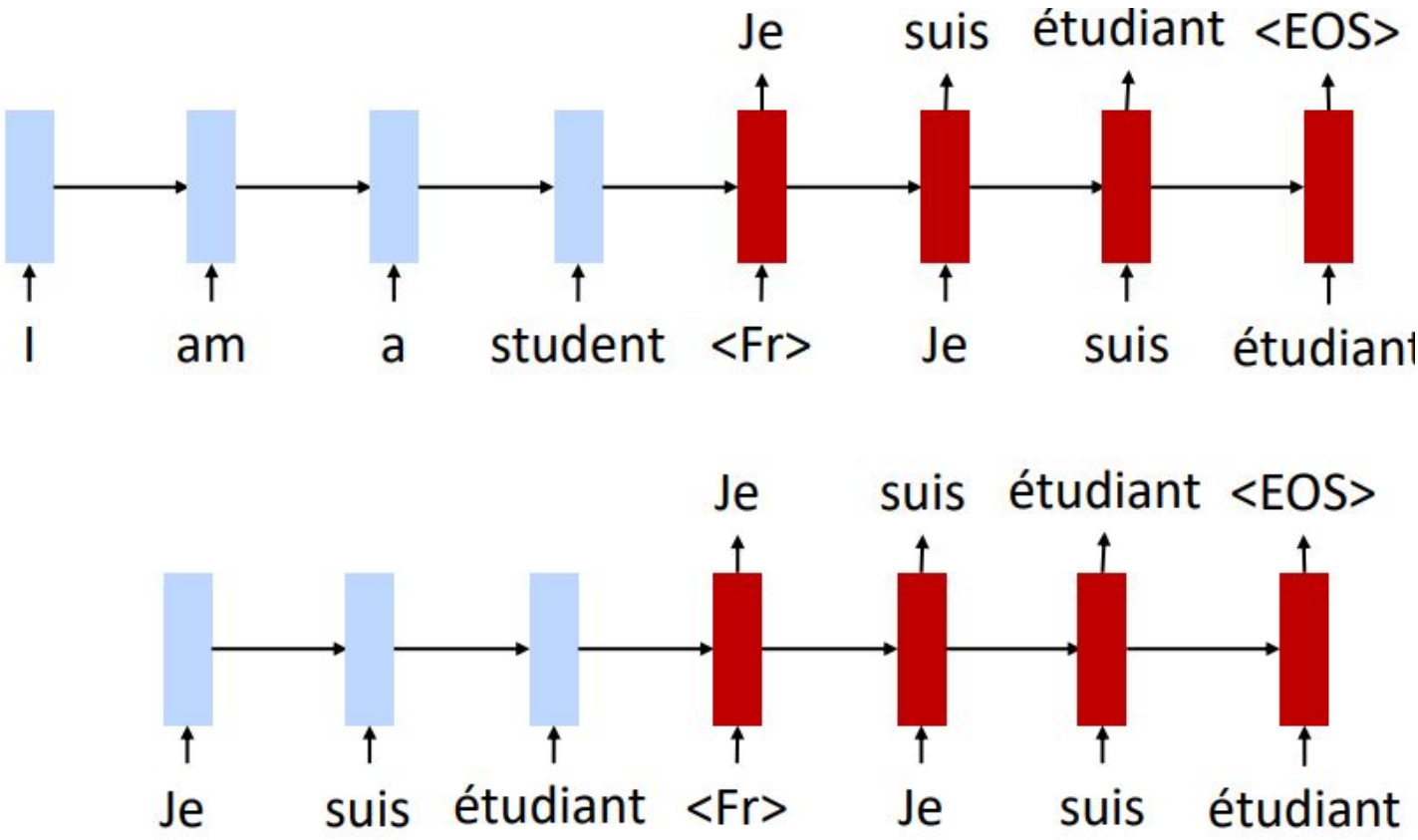
- 引入对抗学习：辨别器用来判断embedding到底是来自Y，还是来自X进行变换的结果（即Wx），训练W，让判别器足够“迷惑”。



无监督词汇翻译

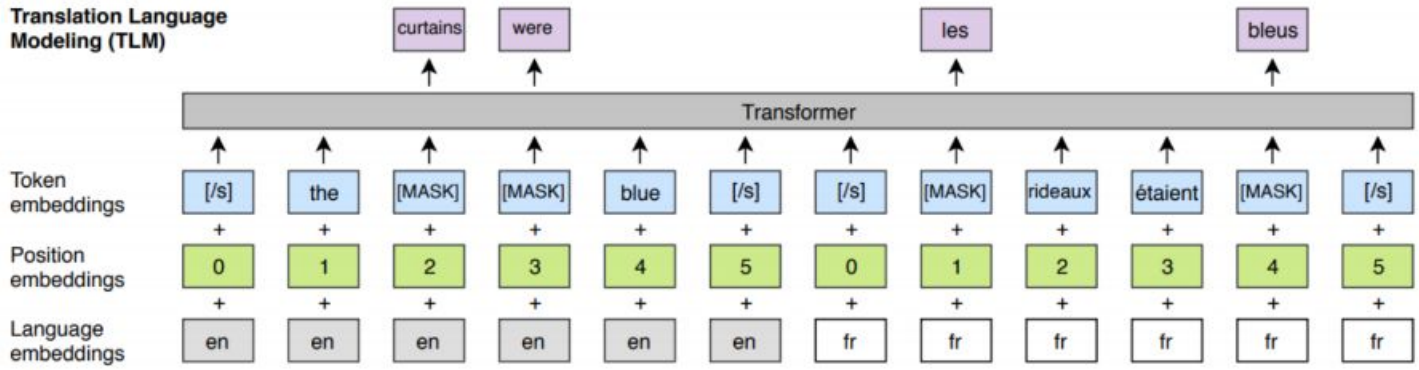
无监督神经网络机器翻译：

- 采用相同的encoder-decoder模型，用于两种语言
- 先执行单个语言内部的自编码（autoencoder）
- 然后进行反向翻译，用反向翻译结果来训练正向翻译



无监督神经网络机器翻译

跨语言BERT



NLP领域的巨型模型

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

机器视觉领域的巨型模型

- 150M参数 [Large Scale GAN Training for High Fidelity Natural Image Synthesis, 2018]
- 550M参数 [GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism, 2018]

专用的模型训练

- 专用硬件、并行训练 [Mesh-TensorFlow: Deep Learning for Supercomputers]

GPT-2模型

- Zero-Shot学习：不需要有监督训练数据
- 阅读理解：<context> <question> A:
- 摘要： <article> TL;DR:
- 翻译：<English sentence1> = <French sentence1>
- 问答：<question> A:

GPT-2公开性争议

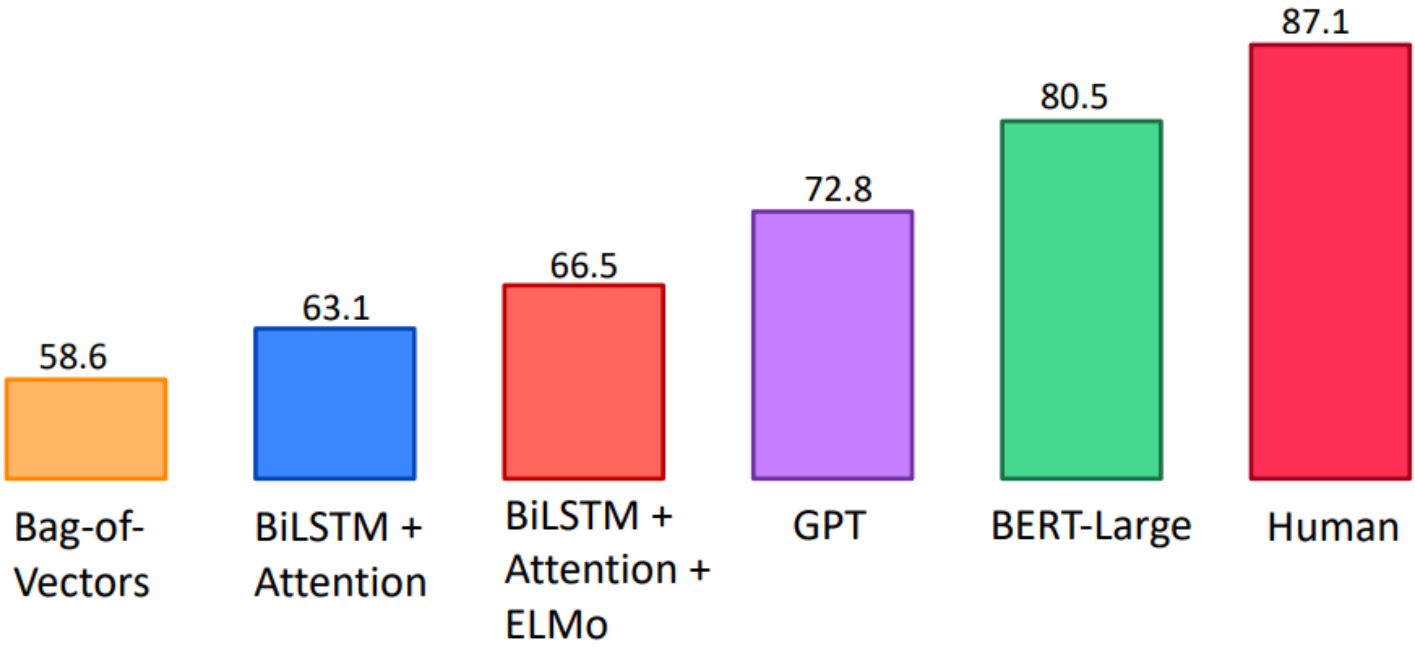
- 官方出于使用考虑不愿意公开完整版本
- 反对者观点。。。
- 支持者观点。。。

更多争议：是否应当让AI从事更高重要性的决定

更多争议：聊天机器人的正面性

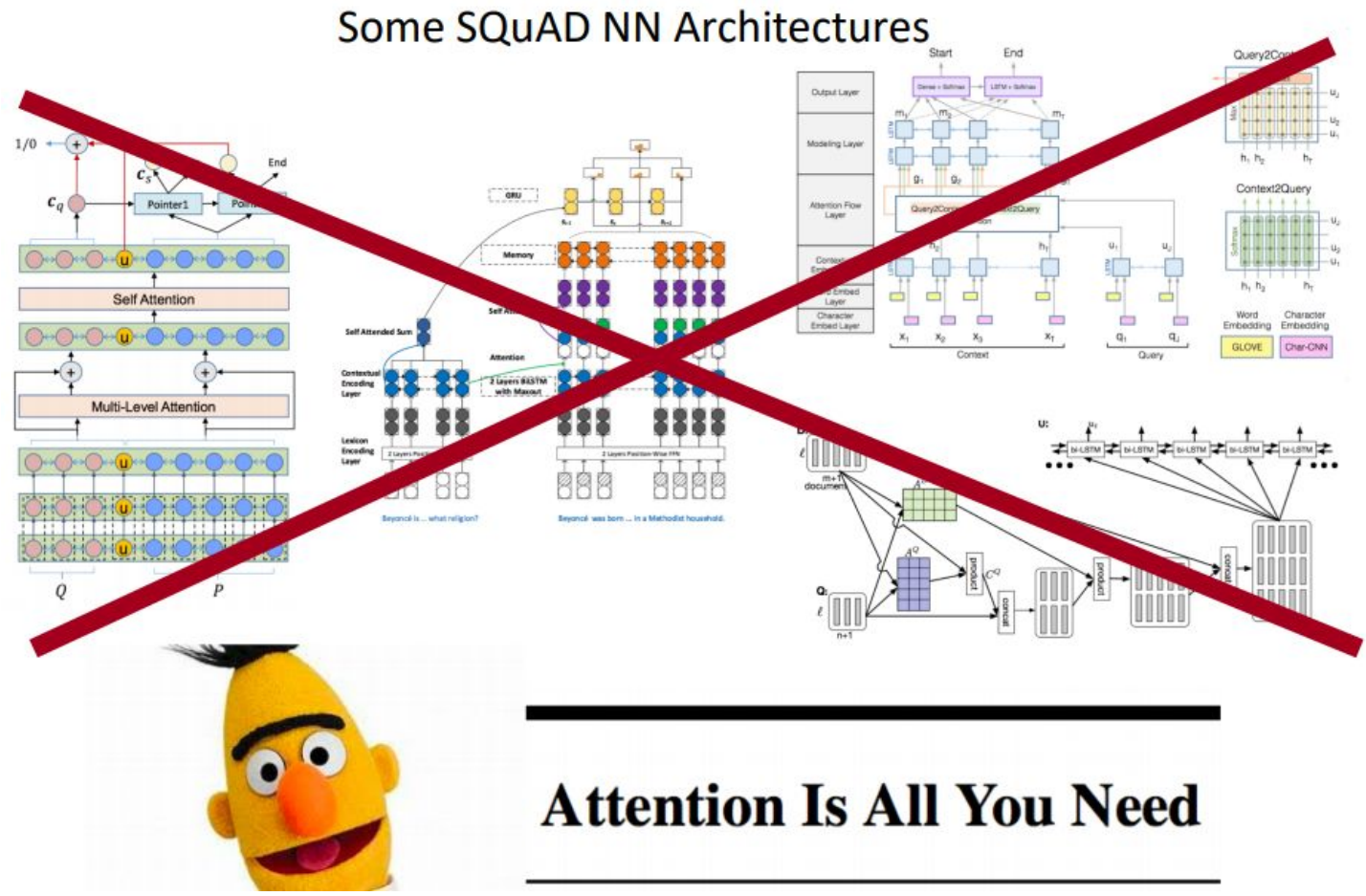
BERT是否解决了所有问题？下一步要做什么？

从GLUE基准测试来看，BERT还没有达到人类水平



是否还需要架构工程？

- 只需要把BERT尺寸扩大几倍，就能获得远比架构工程做几个月高的多的效果提升
- SQuAD上前二十被BERT霸屏



更难的自然语言理解

- 阅读理解：更长文档、多文档、多次跳转的推导、对话中的理解
- 现有的阅读理解数据库的关键问题是：提问时已经看到了上下文，这并不现实，而且会倾向于简单问题
- QuAC: Question Answering in Context
  - 对话之中进行问答，回答问题的老师可以看到维基百科的文章，而提问的学生看不到。
- HotPotQA
  - 需要多次跳转的推理，问题覆盖多篇文档

多任务学习

- GLUE问题
- DecaNLP模型

少资源问题

- 没有足够的算力，比如手机终端
- 少资源的语言
- 少数据，元学习

模型可解释性

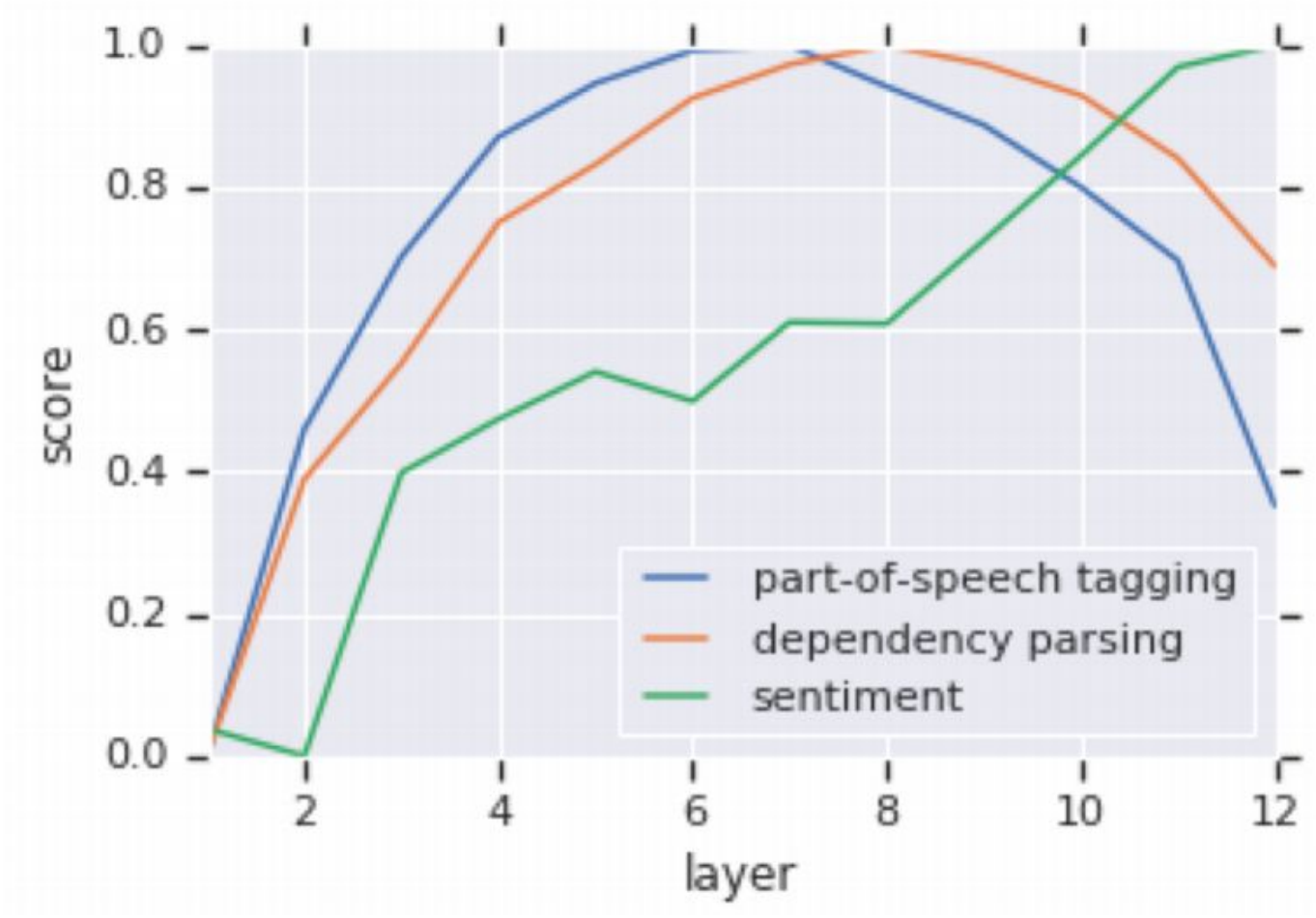
- 是否可以解释模型的预测?
- 是否可以理解BERT之类模型的原理？它们为什么表现很好?

诊断/探索分类

- 用来查看模型到底知道了什么?
- 在模型输出之后接上一些简单任务，用来观察模型的能力
- 常见的诊断任务

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)

常见的诊断任务



对BERT的一次诊断

NLP的工业使用

- 智能音箱
- 聊天机器人
- 客户服务
- 健康分析