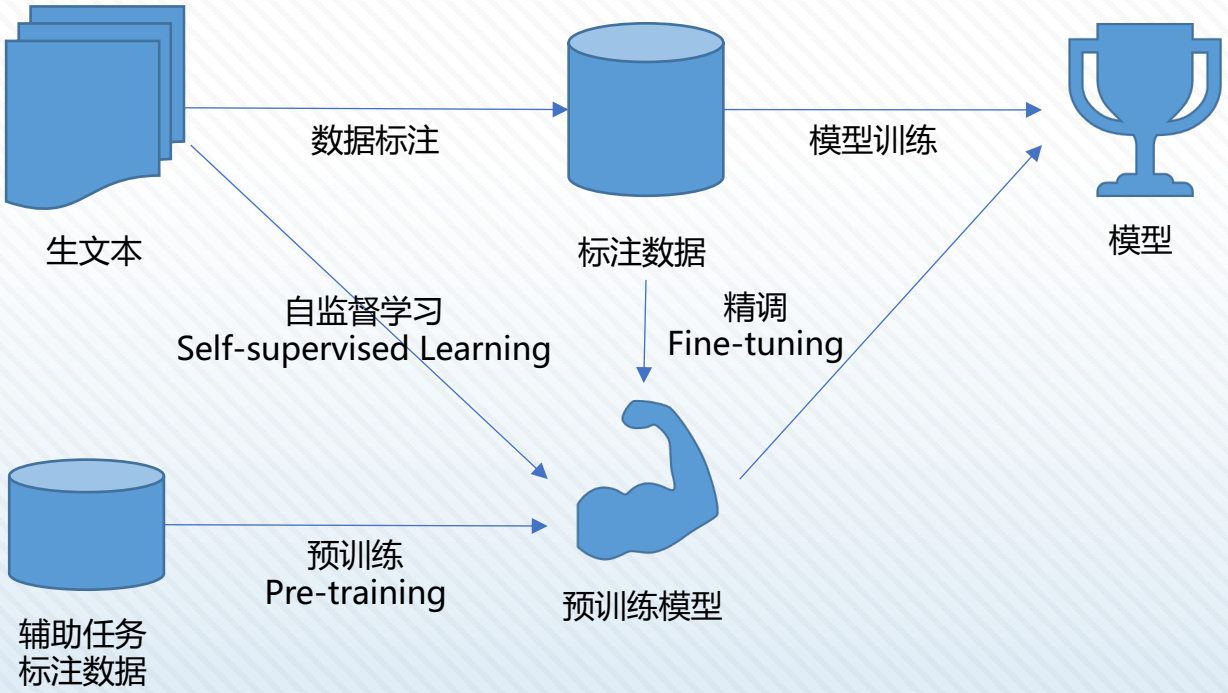
预训练模型–自然语言处理的新范式

车万翔

社会计算与信息检索研究中心
哈尔滨工业大学
2020年12月11日

自然语言处理的发展脉络

预训练模型


数据标注
模型训练
生文本
标注数据
模型
自监督学习
Self-supervised Learning
精调
Fine-tuning
预训练
Pre-training
辅助任务
标注数据
预训练模型

大纲

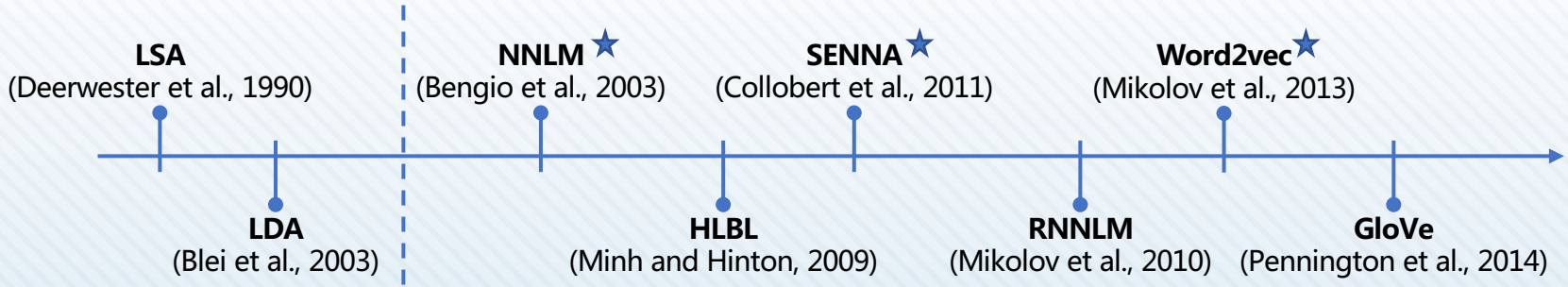
- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的进展
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的进展
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

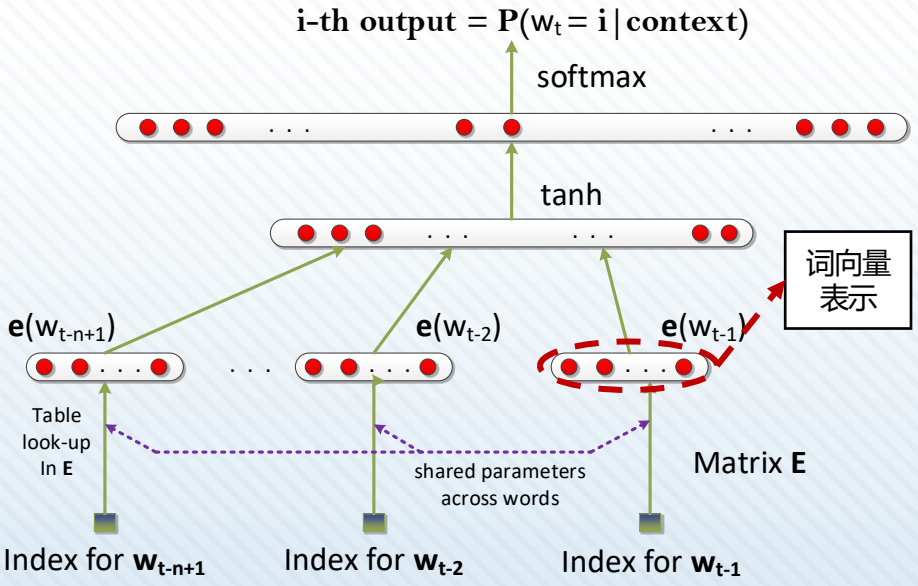
分布式（Distributed）词表示

- 使用低维、稠密、连续的向量表示词
  - 也称词嵌入（Word Embedding）
  - 通过“自指导”的方法直接学习词向量
- 发展历程



神经网络语言模型（NNLM）

- Neural Network Language Models (Bengio et al., JMLR 2003)
  - 根据前n-1个词预测第n个词（语言模型）
  - 模型结构为前向神经网络
  - 通过查表，获得词的向量表示
    - Word Embeddings
    - Word Vectors
  - 通过反向传播优化词向量表示



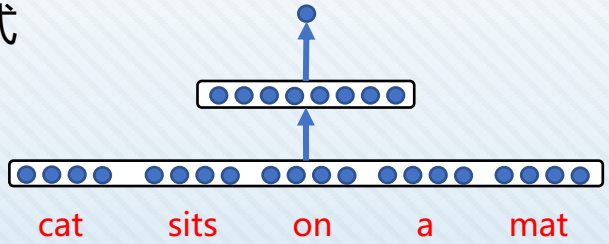
SENNA

- Semantic/syntactic Extraction using a Neural Network Architecture
  - Natural Language Processing (Almost) from Scratch (Collobert et al., JMLR 2011)
- “换词”的思想
  - 一个词和它的上下文构成正例
  - 随机替换掉该词构成负例

✚ cat sits on a mat

▬ cat sits Harbin a mat

- 优化目标
  - *score*(cat sits on a mat) > *score*(cat sits Harbin a mat)
  - *score*的计算方式



- 训练速度慢，在当年的硬件条件下需要训练1个月

Word2vec

- https://code.google.com/archive/p/word2vec/
  (Mikolov et al., ICLR 2013)
- CBOW (Continuous Bag-of-Word)
  - 周围词向量加和预测中间的词
- Skip-Gram
  - 中间词预测周围词
- 训练速度快
  - 线性模型
  - 可利用大规模数据
  - 弥补了模型能力的不足

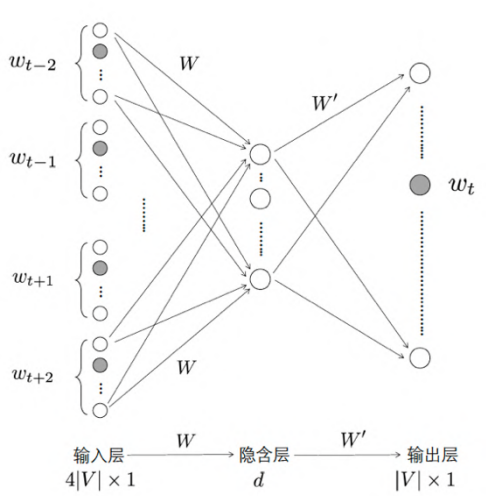


Figure 3: CBOW模型。

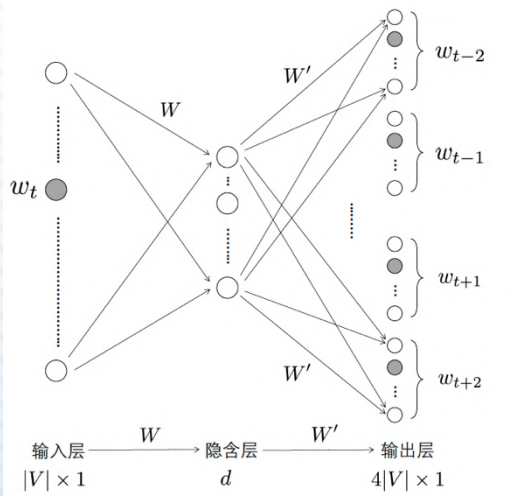
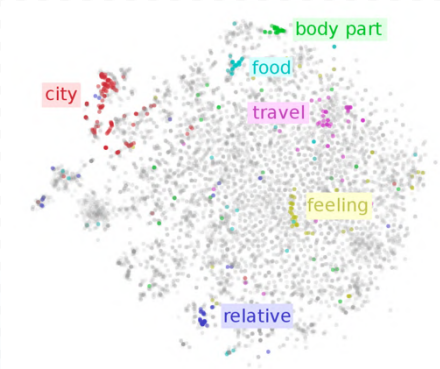
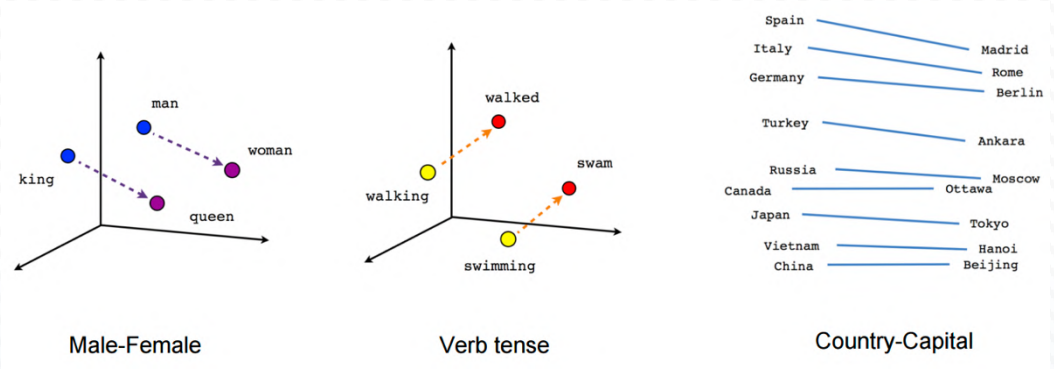


Figure 4: Skip-gram模型。

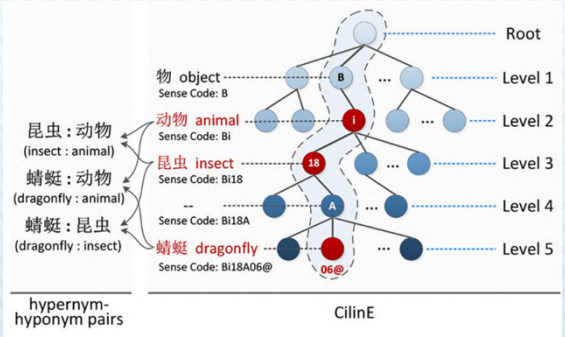
词向量的应用



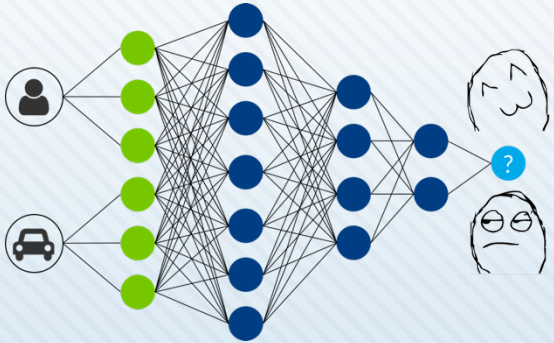
词义相似度计算



词类比关系计算



知识图谱补全



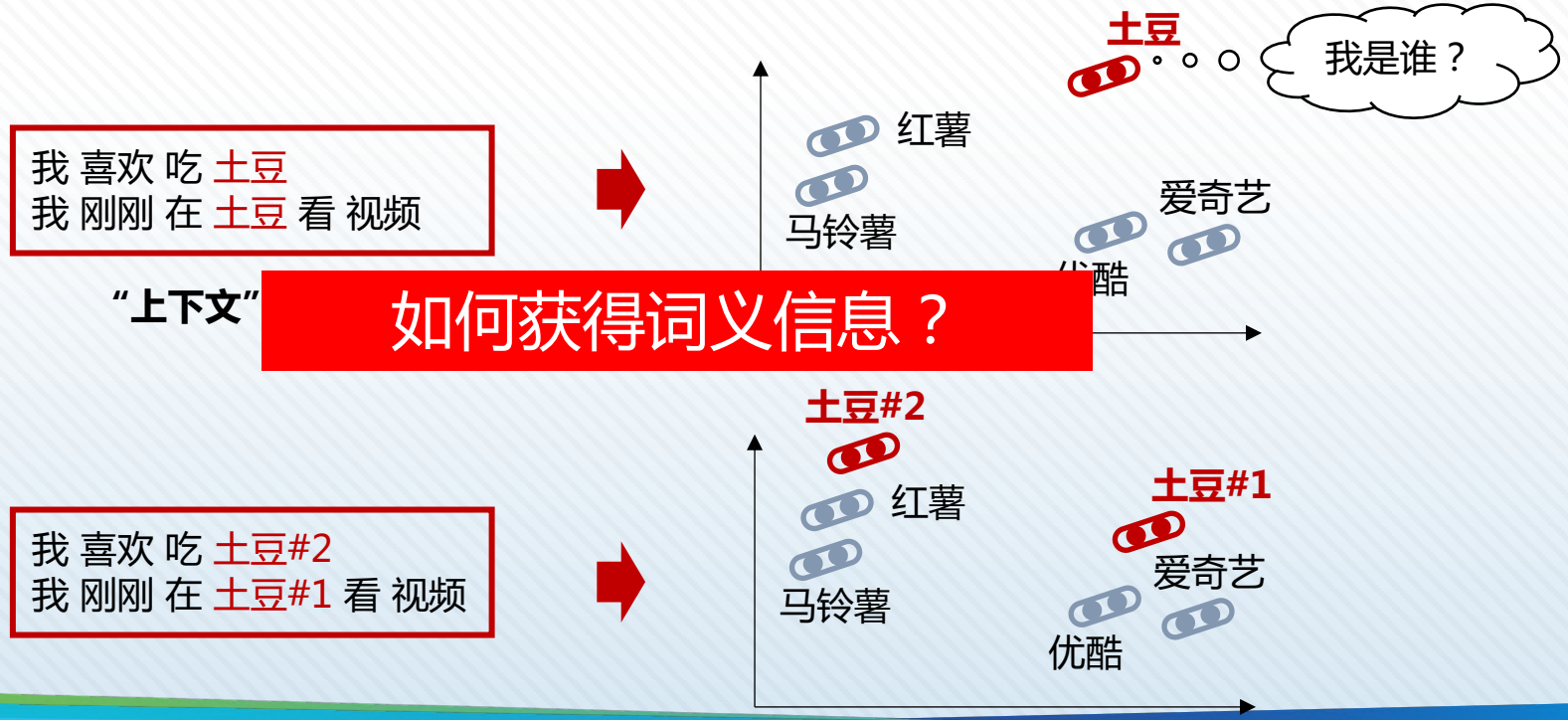
推荐系统

大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的进展
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

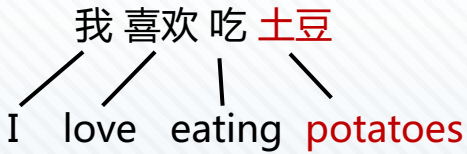
一词多义现象

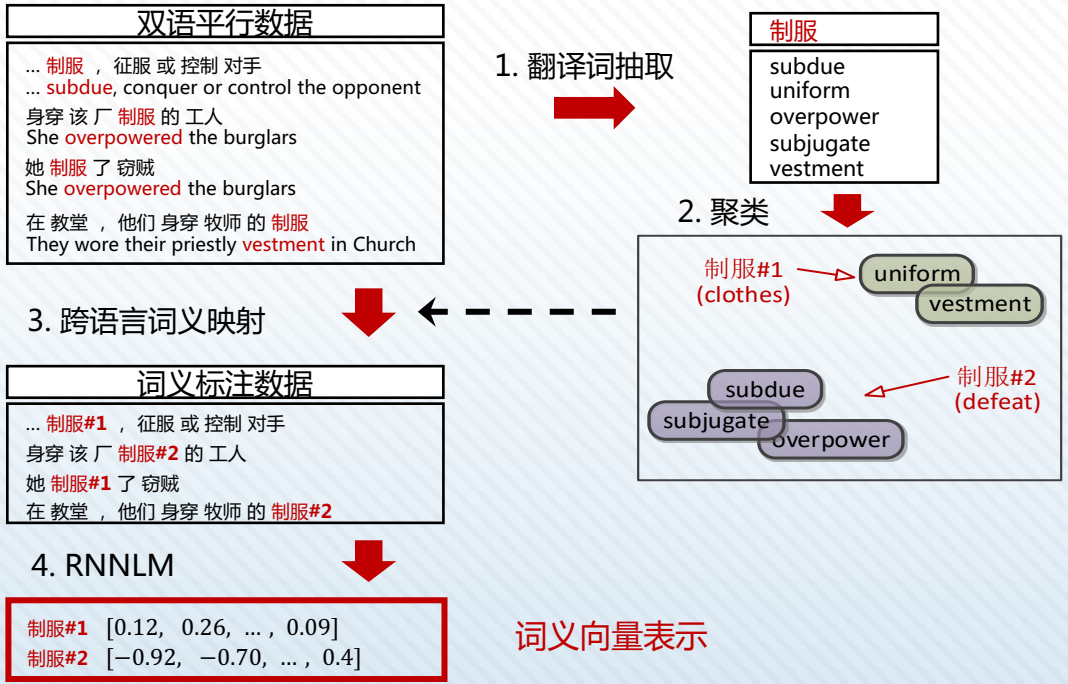
- 以上所有工作都假设一个词由唯一的词向量表示
  - 无法处理一词多义现象



基于双语的词义向量表示

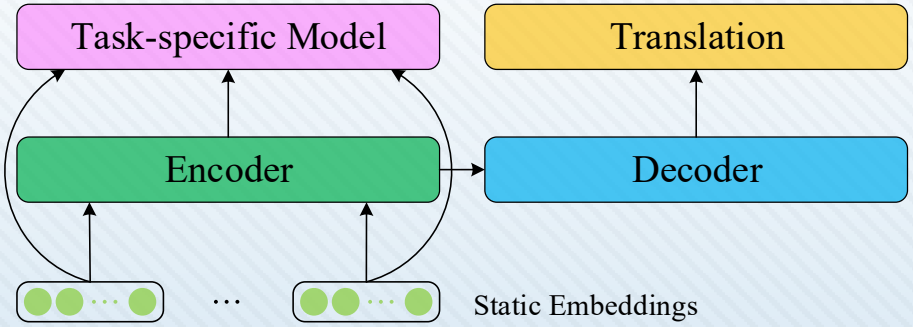
- Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources (Guo et al., Coling 2014[SCIR])

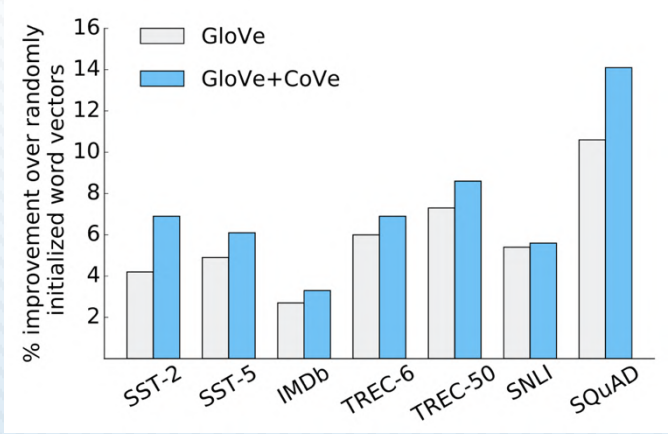




CoVe

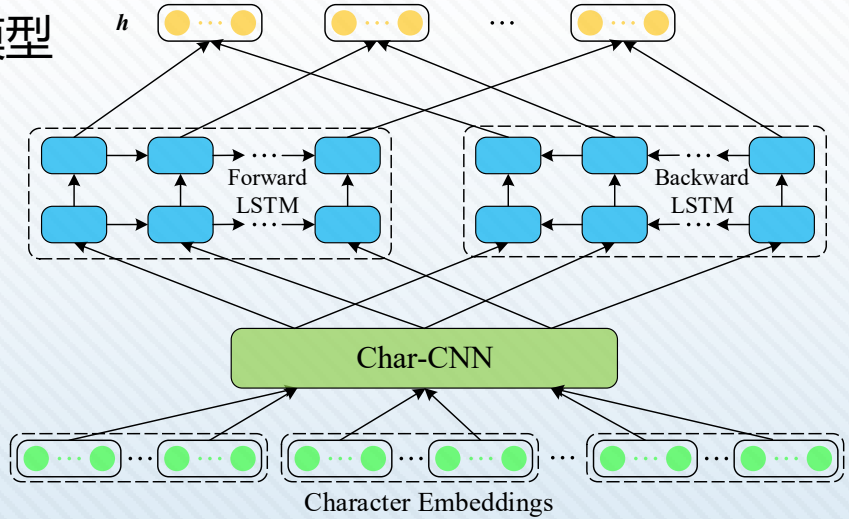
- Learned in Translation: Contextualized Word Vectors (McCann et al., arXiv:1708.00107)
  - CoVe: Context Vectors
- 预训练NMT模型
- 将Encoder作为目标任务的额外特征





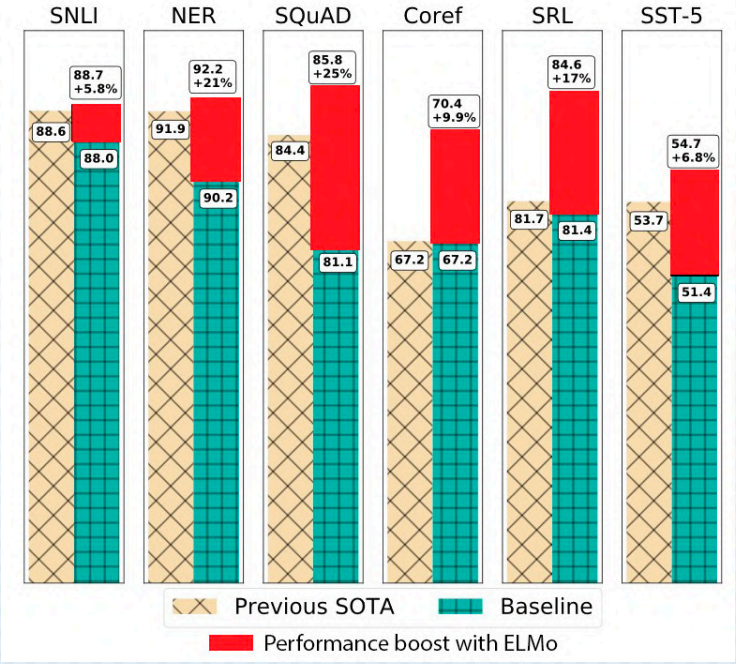
ELMo

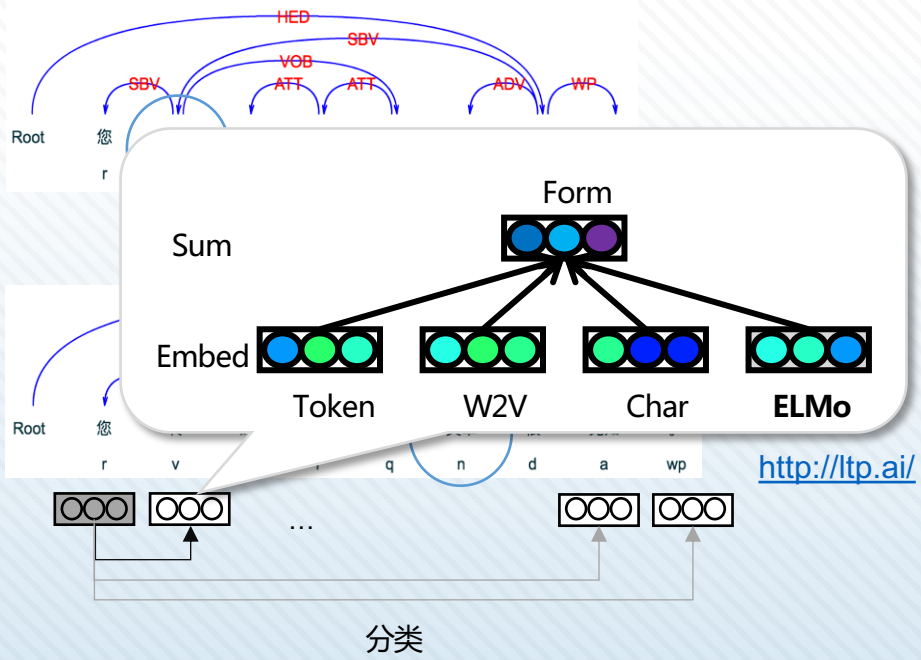
- Deep Contextualized Word Representations (Peters et al., NAACL 2018 Best Paper)
  - ELMo: Embeddings from Language Models
- 使用字符的CNN表示词
- 分别训练从左至右和从右至左的语言模型
- 使用语言模型的输出作为词向量特征
- 语言模型训练数据接近“无限”



基于ELMo的应用

- 依存句法分析 (Che et al., CoNLL 2018[SCIR])



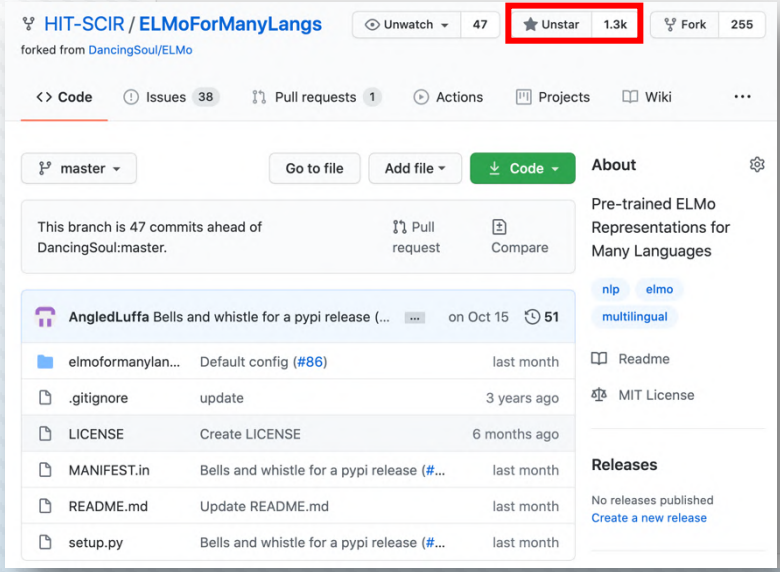


CoNLL 2018评测

- http://universaldependencies.org/conll18/
- Multilingual Parsing from Raw Text to Universal Dependencies
  - 包括分句、分词、词性标注、依存句法分析任务
  - 数据：57种语言、82个树库
- 技术方案
  - **ELMo**、集成学习、多树库融合
- 哈工大获得第1名，高出第2名2.5%
- 多国语ELMo开源
  - https://github.com/HIT-SCIR/ELMoForManyLangs

LAS Ranking

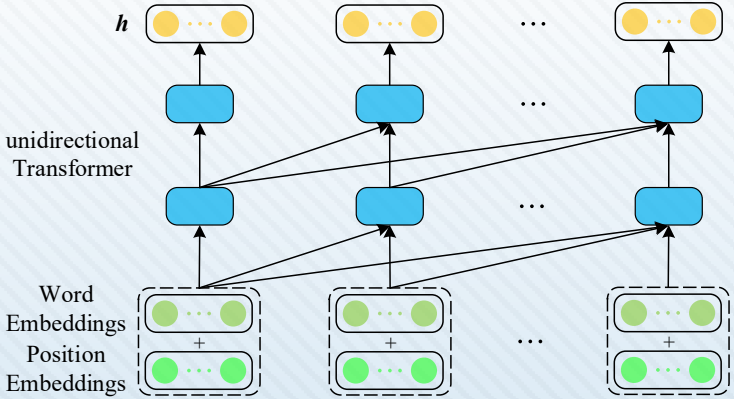
1.	HIT-SCIR (Harbin)	75.84 ± 0.14 [OK] (p<0.001)
2.	TurkuNLP (Turku)	73.28 ± 0.14 [OK] (p=0.039)
3-5.	UDPipe Future (Praha)	73.11 ± 0.13 [OK] (p=0.221)
3-5.	LATTICE (Paris)	73.02 ± 0.14 [OK] (p=0.461)
3-5.	ICS PAS (Warszawa)	73.02 ± 0.14 [OK] (p<0.001)
6.	CEA LIST (Paris)	72.56 ± 0.14 [OK] (p=0.036)
		(p=0.191)
		(p<0.001)
		(p=0.242)
		(p=0.032)
		(p<0.001)
		(p<0.001)
		(p<0.001)
		(p=0.002)
		(p<0.001)
		(p=0.015)
		(p<0.001)
		(p=0.048)
		(p<0.001)
		(p<0.001)
		(p<0.001)
		(p<0.001)
		(p<0.001)
		(p<0.001)
		(p<0.001)

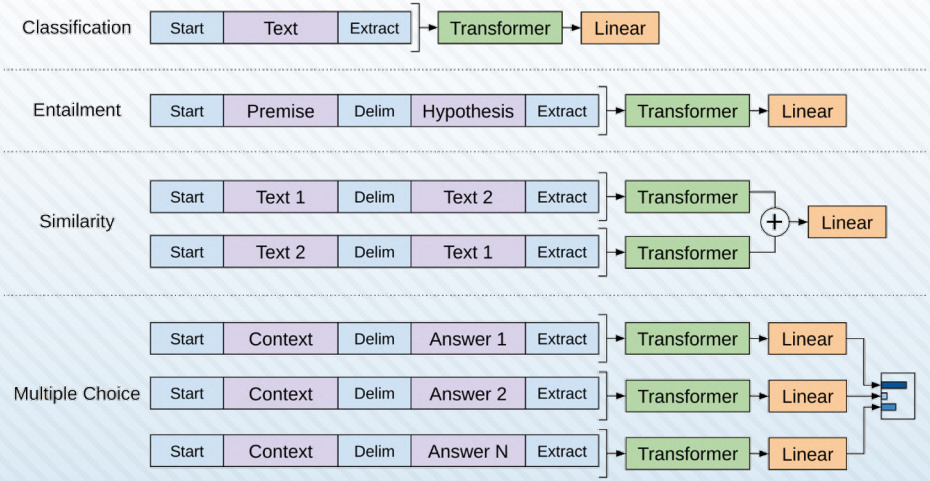


大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的进展
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

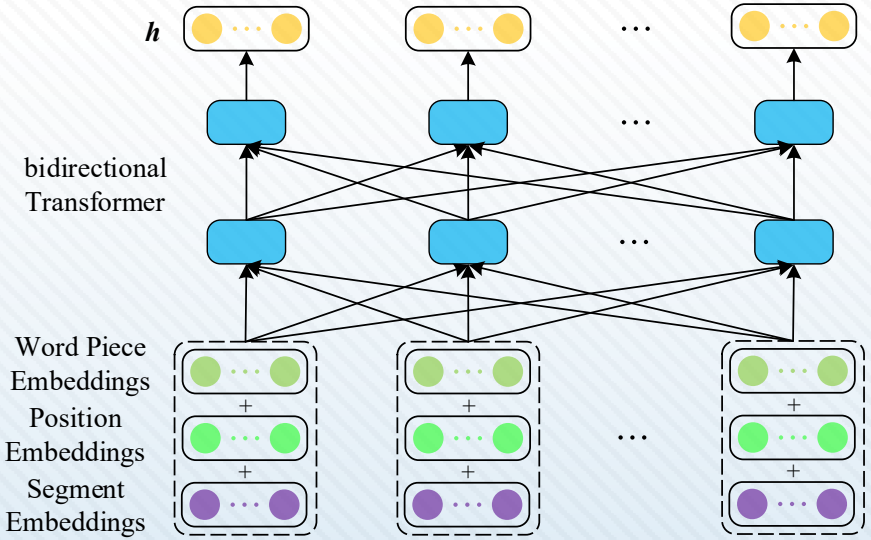
- Improving Language Understanding by Generative Pre-Training (Radford et al., 2018)
  - GPT: Generative Pretrained Transformer
  - 使用12层的Transformer作为Encoder预训练单向语言模型
  - 在目标任务上精调（Fine-tuning）模型





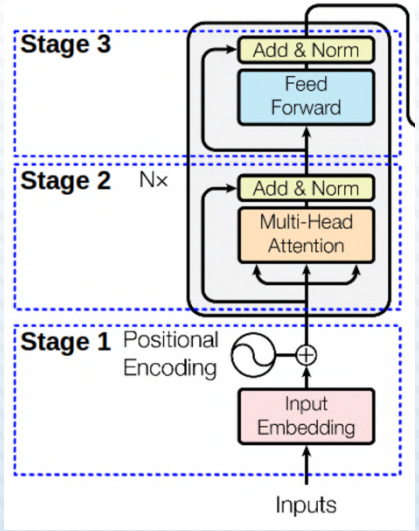
BERT

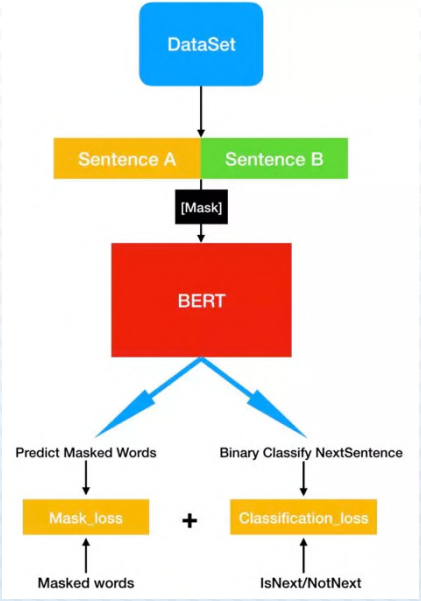
- Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019 Best Paper)
  - BERT: **Bidirectional** Encoder Representations from Transformers

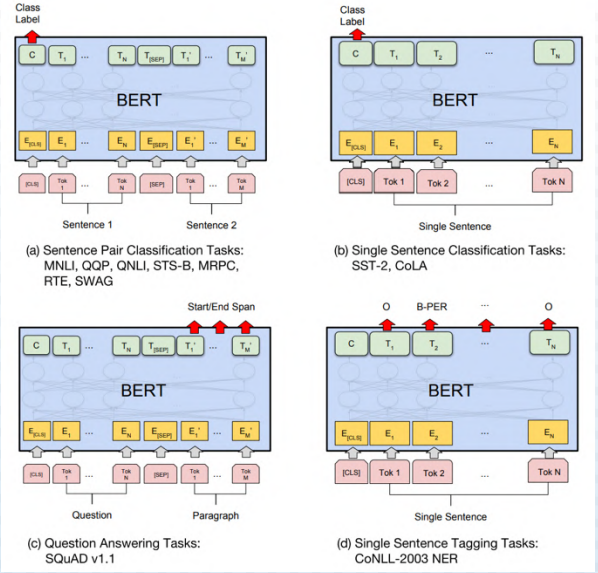


BERT模型详解

- 编码器
  - 输入：Word Piece
  - 编码器：Transformer
- 预训练任务
  - 完形填空 + 下句预测 (NSP)
- 应用方式
  - 在目标任务上Fine-tune
  - 四种任务类型

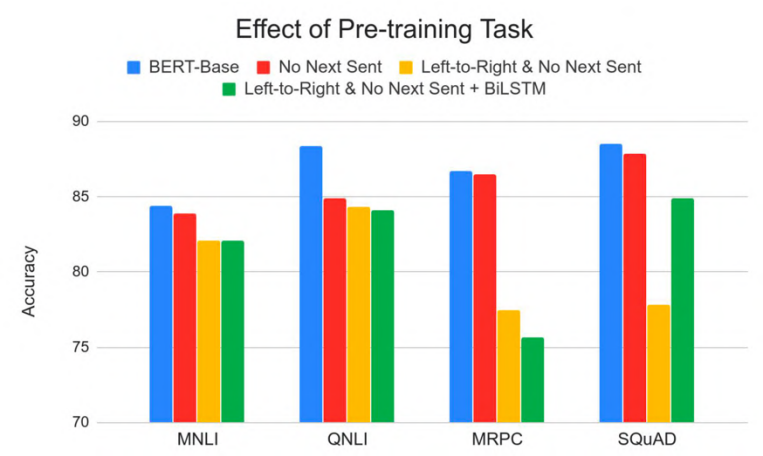




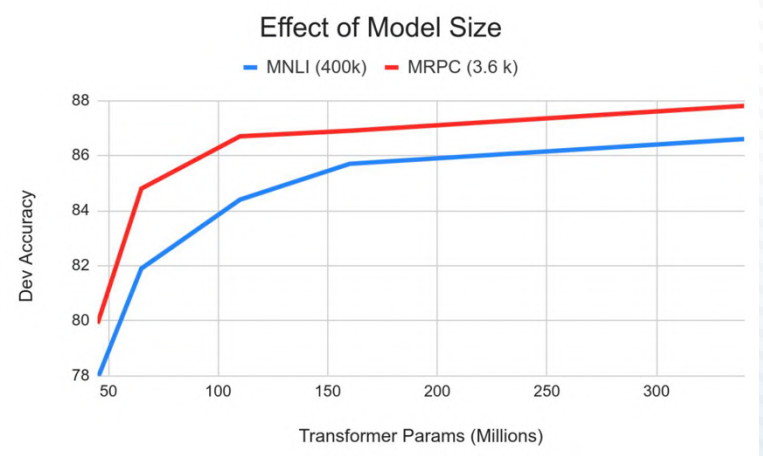


BERT中各种策略的影响

预训练任务



模型大小



大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的进展
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

预训练模型的进展

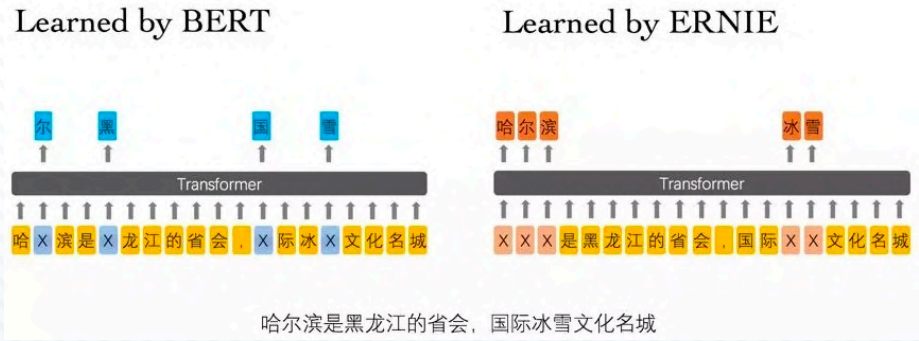
- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

ERNIE（百度）

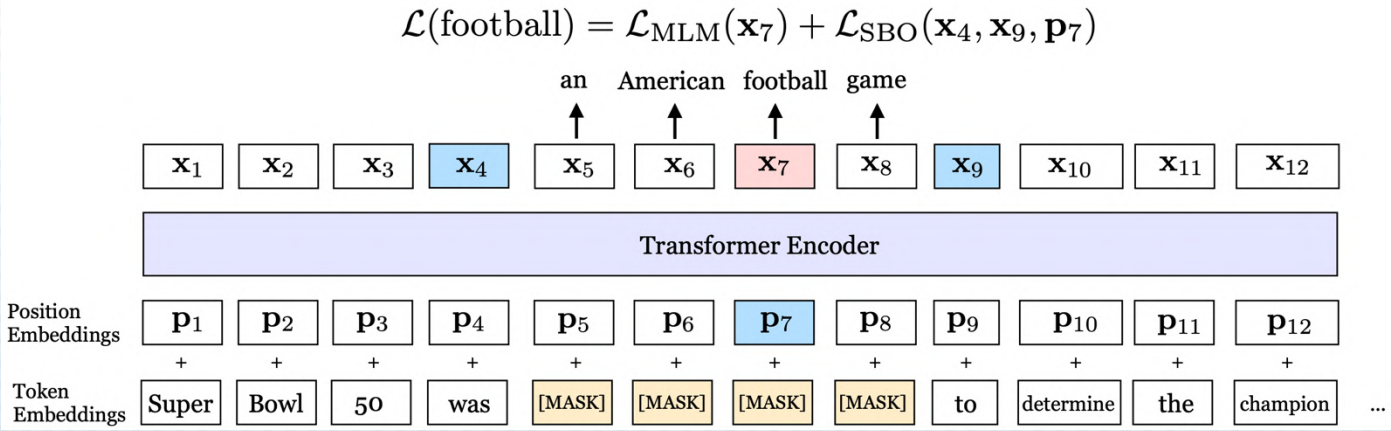
- Enhanced Representation through Knowledge Integration (Sun et al., arXiv:1904.09223, AAAI 2020)
  - ERNIE 1.0
    - Mask词和实体
  - ERNIE 2.0
    - 更多的预训练任务
    - 更丰富的预训练数据



任务	ERNIE 1.0 模型	ERNIE 2.0 英文模型	ERNIE 2.0 中文模型
Word-aware	Knowledge Masking	Knowledge Masking Capitalization Prediction Token-Document Relation Prediction	Knowledge Masking
Structure-aware		Sentence Reordering	Sentence Reordering Sentence Distance
Semantic-aware	Next Sentence Prediction	Discourse Relation	Discourse Relation IR Relevance

SpanBERT

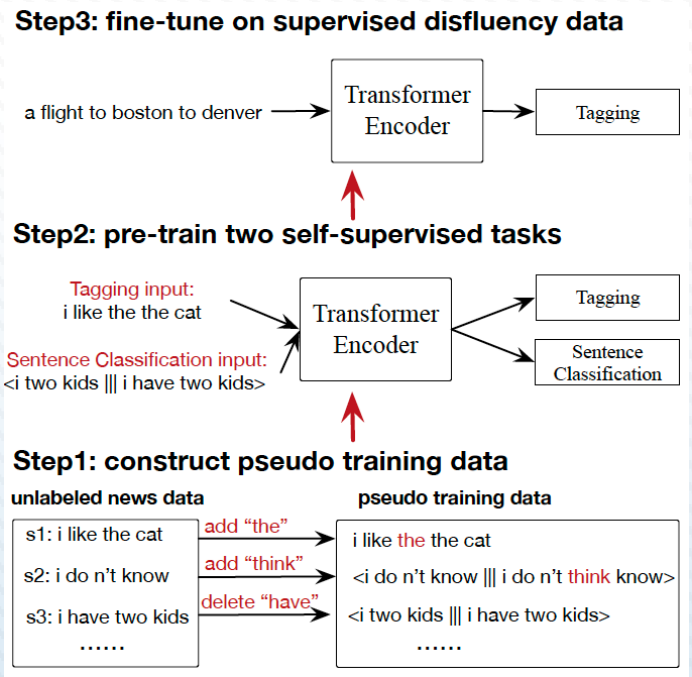
- SpanBERT: Improving Pre-training by Representing and Predicting Spans (Joshi et al., arXiv:1907.10529)
  - 挖掉一段文字，通过学习段的边界表示预测段中每个词
  - 去除NSP预训练目标（由于主题不同，容易判断）
  - 在段抽取任务，如抽取式问答中表现良好



伪顺滑数据预训练

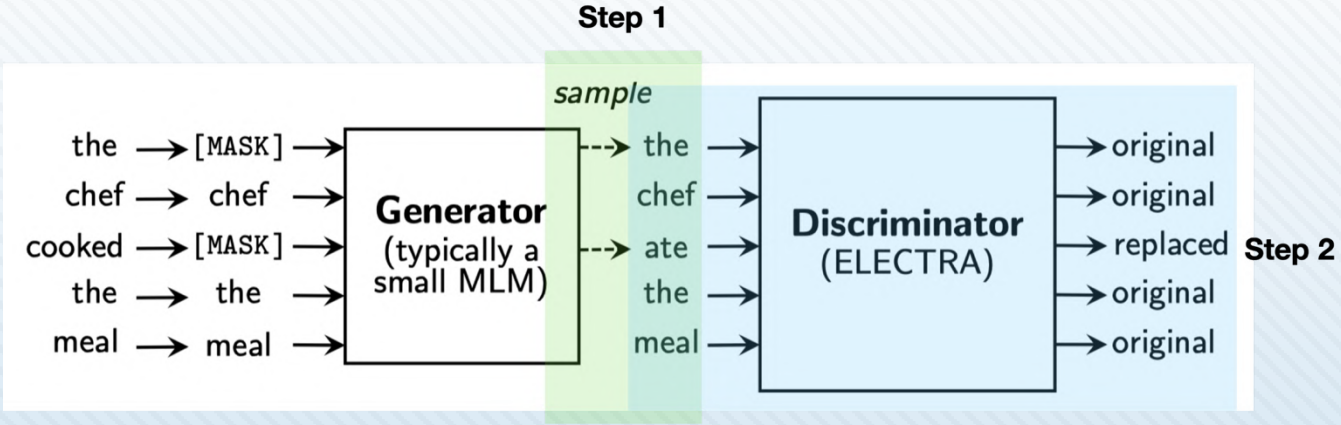
- Multi-Task Self-Supervised Learning for Disfluency Detection (Wang et al., AAAI 2020[SCIR])
  - 随机增删改原句中的词，构成非顺滑数据
  - 两个预训练任务
    - 预测增删改的词
    - 判断哪个句子是原句
  - 显著提升顺滑任务的准确率

Method	P	R	F1
UBT (Wu et al. 2015)	90.3	80.5	85.1
Semi-CRF (Ferguson et al., 2015)	90.0	81.2	85.4
Bi-LSTM (Zayats et al., 2016)	91.8	80.6	85.9
LSTM-NCM (Lou and Johnson 2017)	-	-	86.8
Transition-based (Wang et al. 2017)	91.1	84.1	87.5
Our self-supervised (1000 sents)	88.6	83.7	86.1
Our self-supervised (Full)	**93.4**	**87.3**	**90.2**



ELECTRA

- ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (Clark et al., ICLR 2020)
  - 通过BERT采样生成数据
  - 与原始数据比较，生成序列级训练样本
  - 预训练数据增多，模型收敛更快
  - 预训练数据与真实数据相似，模型效果提升



预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

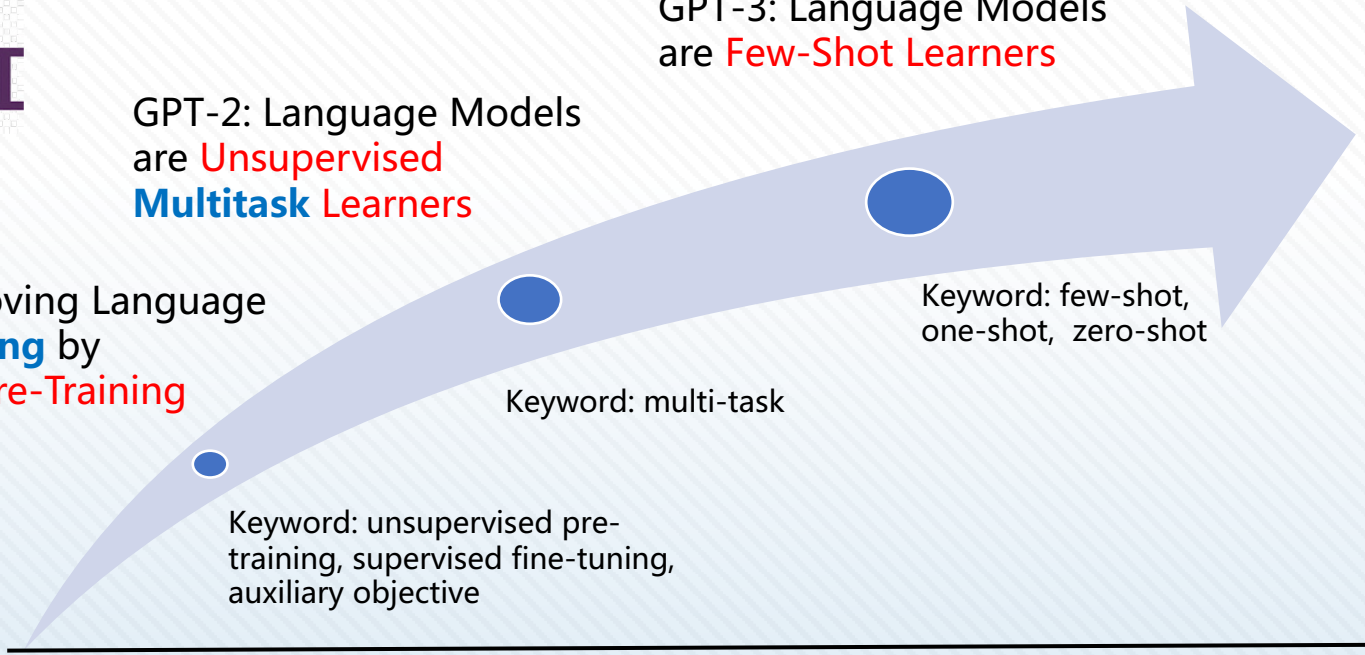
GPT系列模型

OpenAI

GPT-1: Improving Language **Understanding** by Generative Pre-Training

GPT-2: Language Models are Unsupervised **Multitask** Learners

GPT-3: Language Models are Few-Shot Learners



GPT系列模型对比

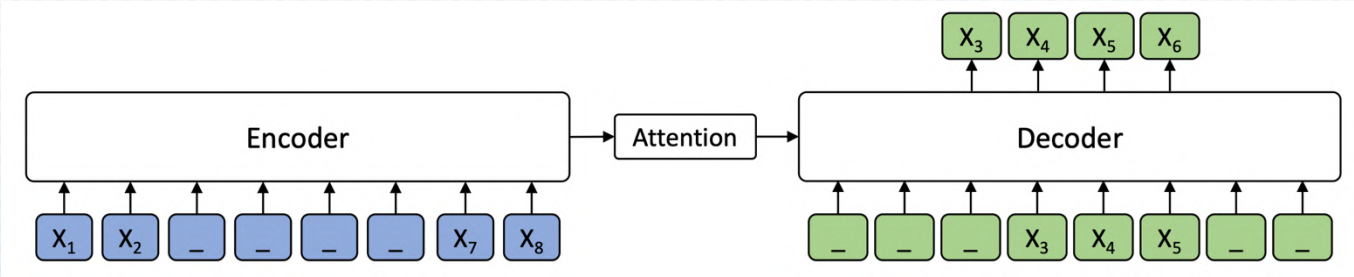
模型规模对比

模型	规模
GPT	$d_{model} = 768, context\_size = 512, layer\_num = 12$
GPT-2	$d_{model} = 1600, context\_size = 1024, layer\_num = 48, param = 1.5B, size = 774M$
GPT-3	$d_{model} = 12288, context\_size = 2048, layer\_num = 96, param = 175B, size = 70G$

数据规模对比

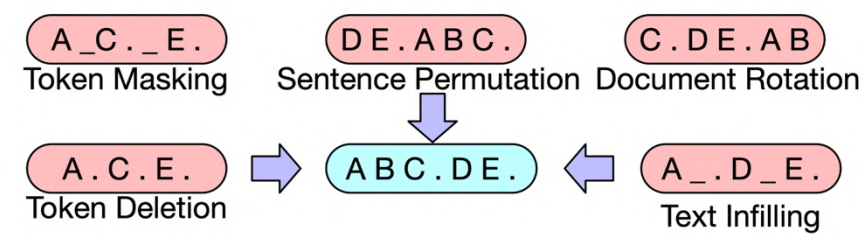
MASS

- MASS: Masked Sequence to Sequence Pre-training for Language Generation (Song et al., arXiv:1905.02450)
  - 挖掉句子中的一段文字
  - 通过其余部分，使用seq2seq模型重构该段文字
  - 更适应于语言生成任务，如神经机器翻译

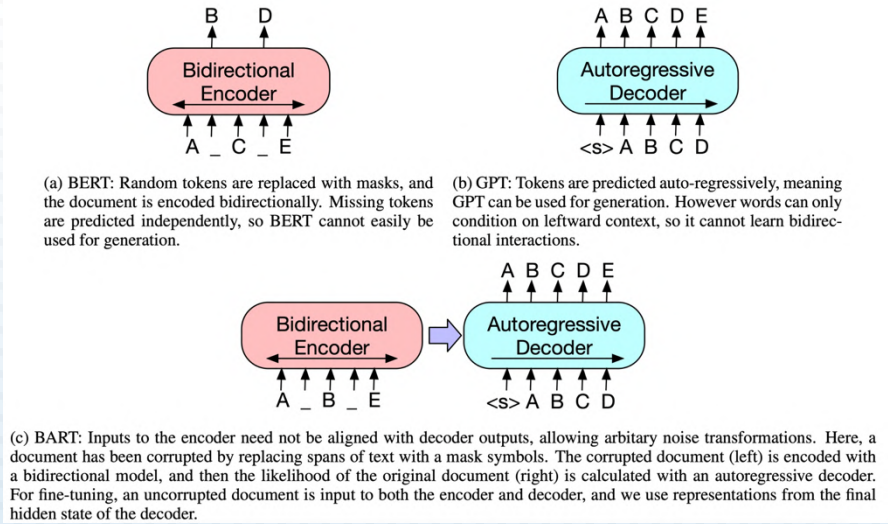


BART

- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (Lewis et al., ACL 2020)
  - 2个训练步骤
    - 使用任意噪声破坏文本
    - 模型学习重建原始文本
  - 在文本生成类任务上表现良好



噪声函数



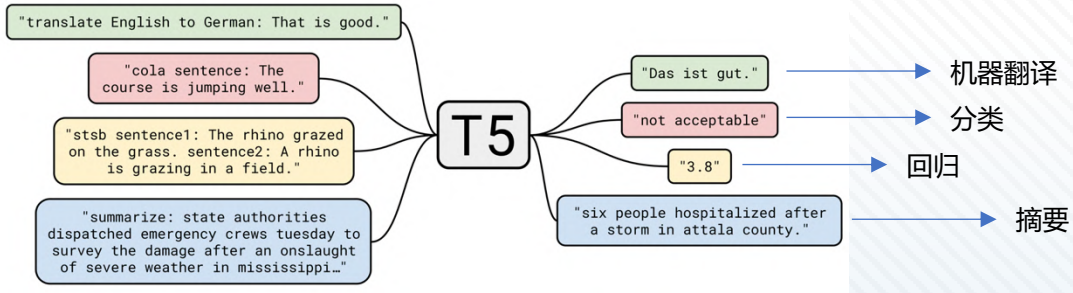
(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

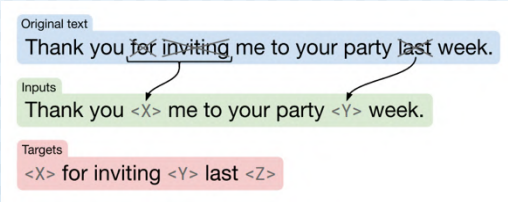
(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

T5

- T5: Exploring the Limits of Transfer Learning with a Unified **T**ext-**t**o-**T**ext **T**ransformer (Raffel et al., arXiv:1910.10683)
  - 把所有的NLP问题都可以定义成“text-to-text”问题



  - 约750G的清洗后的C4（Colossal Clean Crawled Corpus）语料库
  - 训练步骤



预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

RoBERTa

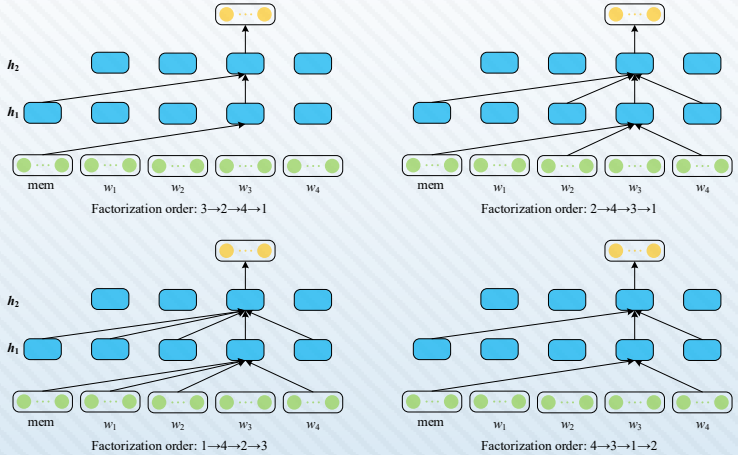
- RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., arXiv:1907.11692)
  - 基于BERT进行细致调参
    - 更多的数据
    - 更多的迭代步数
    - 更大的batch
      - 256 → 8192
    - 更大的BPE词表
      - 30K → 50K
    - 去除NSP任务
    - 训练过程中，动态改变Mask的内容
  - 在1,024块V100 GPU上训练了一天！！

预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

XLNet

- XLNet: Generalized Autoregressive Pretraining for Language Understanding (Yang et al., arXiv:1906.08237)
  - 使用Transformer-XL对长序列建模 (Dai et al., ACL 2019)
  - 已有模型的问题
    - 自回归语言模型（根据上文预测下一个词）看不到下文
    - 自编码语言模型（根据上下文预测中间的内容）预训练和精调时输入不一致
  - 解决方案
    - 随机排列各种词序输入自回归语言模型
    - 解决看不到下文的问题

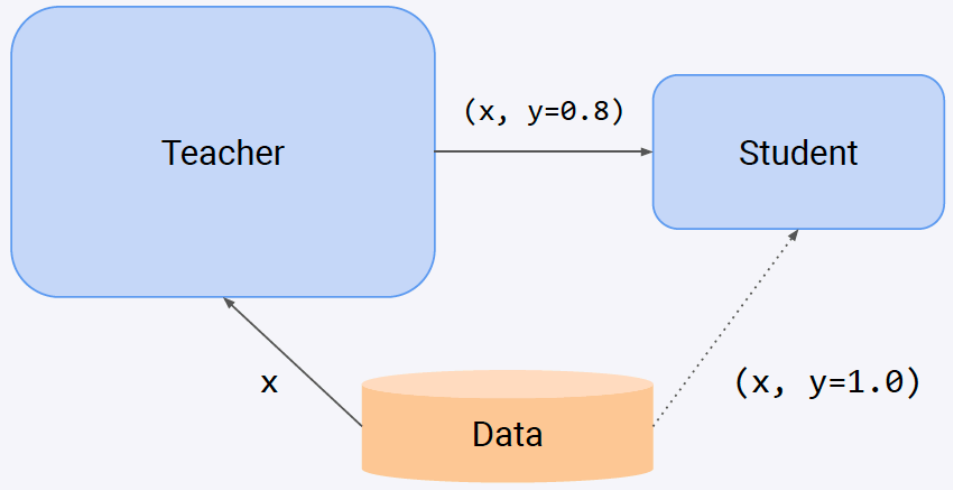


预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

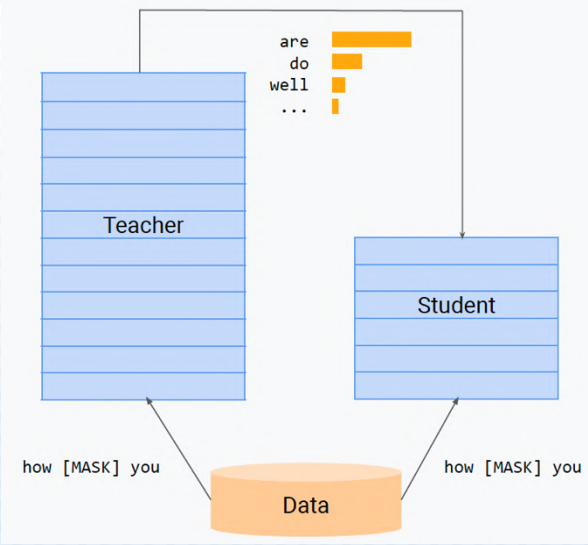
知识蒸馏（Knowledge Distillation）

- Distilling the Knowledge in a Neural Network (Hinton et al., arXiv:1503.02531)
- 使用小模型（Student）模仿大模型（Teacher）的预测结果
- 将大模型的知识迁移到小模型中
- 无明显的性能损失



DistilBERT

- Distilling BERT (Sanh et al., NeurIPS Workshop 2019)
  - 蒸馏：使用小模型，模仿大模型的预测结果



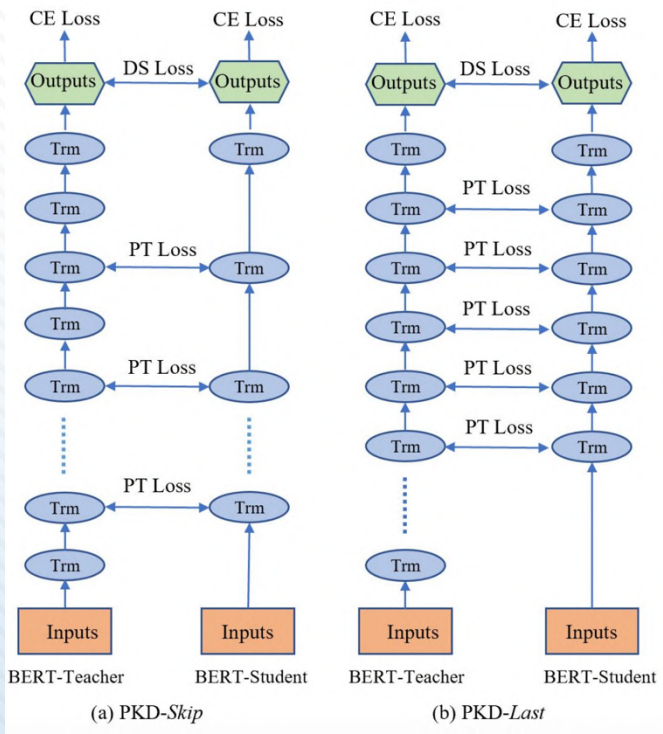
	Nb of parameters (millions)	Inference Time (s)
GLUE BASELINE (ELMo + BiLSTMs)	180	895
BERT base	110	668
DistilBERT	66	410

	Macro Score	CoLA	MNLI	MNLI-MM	MRPC		QNLI	QQP		RTE	SST-2	STS-B		WNLI
		mcc	acc	acc	acc	f1	acc	acc	f1	acc	acc	pearson	spearmanr	acc
GLUE BASELINE (ELMo + BiLSTMs)	68.7	44.1	68.6 (avg)		70.8	82.3	71.1	88.0	84.3	53.4	91.5	70.3	70.5	56.3
BERT base	78.0	55.8	83.7	84.1	86.3	90.5	91.1	90.9	87.7	68.6	92.1	89.0	88.6	43.7
DistilBERT	75.2	42.5	81.6	81.1	82.4	88.3	85.5	90.6	87.7	60.0	92.7	84.5	85.0	55.6

PKD for BERT

- Patient Knowledge Distillation for BERT Model Compression (Sun et al., EMNLP 2019)
  - 按层蒸馏：不只模拟输出层
  - 跳层蒸馏：进一步减小参数量

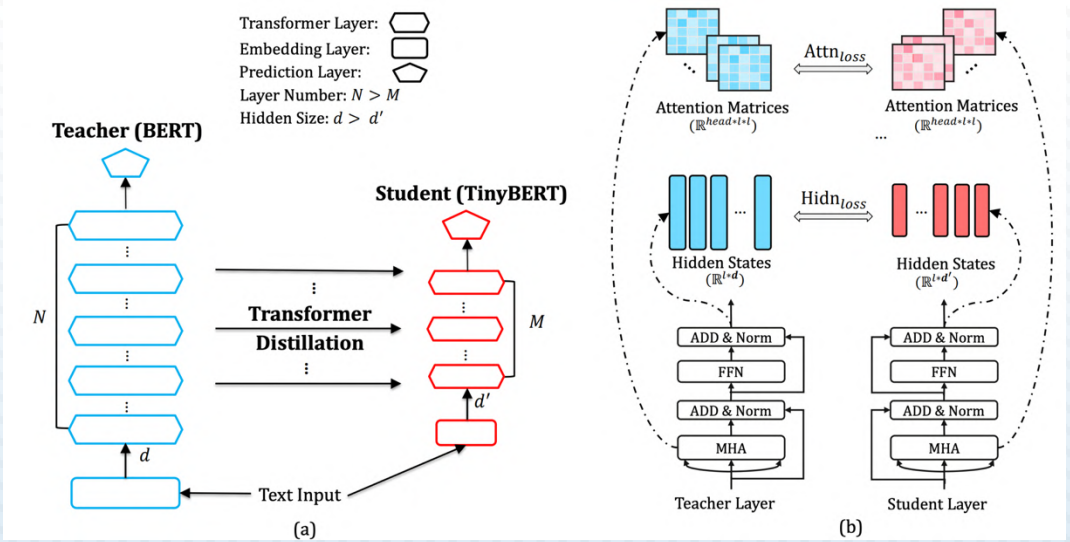
Model	SST-2 (67k)	MRPC (3.7k)	QQP (364k)	MNLI-m (393k)	MNLI-mm (393k)	QNLI (105k)	RTE (2.5k)
$BERT_{12}$ (Google)	93.5	88.9/84.8	71.2/89.2	84.6	83.4	90.5	66.4
$BERT_{12}$ (Teacher)	94.3	89.2/85.2	70.9/89.0	83.7	82.8	90.4	69.1
$BERT_6$-FT	90.7	85.9/80.2	69.2/88.2	80.4	79.7	86.7	63.6
$BERT_6$-KD	91.5	**86.2/80.6**	70.1/88.8	80.2	79.8	88.3	64.7
$BERT_6$-PKD	**92.0**	85.0/79.9	**70.7/88.9**	**81.5**	**81.0**	**89.0**	**65.5**
$BERT_3$-FT	86.4	80.5/**72.6**	65.8/86.9	74.8	74.3	84.3	55.2
$BERT_3$-KD	86.9	79.5/71.1	67.3/87.6	75.4	74.8	84.0	56.2
$BERT_3$-PKD	**87.5**	**80.7**/72.5	**68.1/87.8**	**76.7**	**76.3**	**84.7**	**58.2**



TinyBERT

- TinyBERT: Distilling BERT for Natural Language Understanding (Jiao et al., arXiv:1909.10351)
  - 学习目标模型（Teacher）的
    - 隐层激活
    - 注意力矩阵
  - 最高压缩7.5倍
  - 推理速度快9.4倍

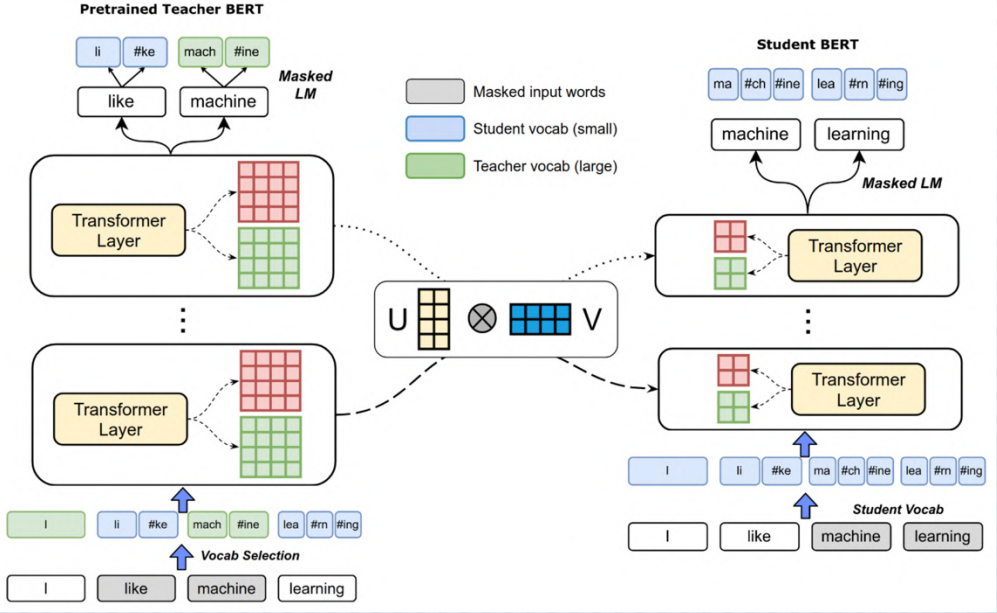
System	Average
$\text{BERT}_{\text{BASE}}$ (Google)	79.6
$\text{BERT}_{\text{BASE}}$ (Teacher)	79.5
$\text{BERT}_{\text{SMALL}}$	70.2
Distilled $\text{BiLSTM}_{\text{SOFT}}$	-
BERT-PKD	72.6
DistilBERT	71.9
TinyBERT	76.5



词表优化与逐层映射

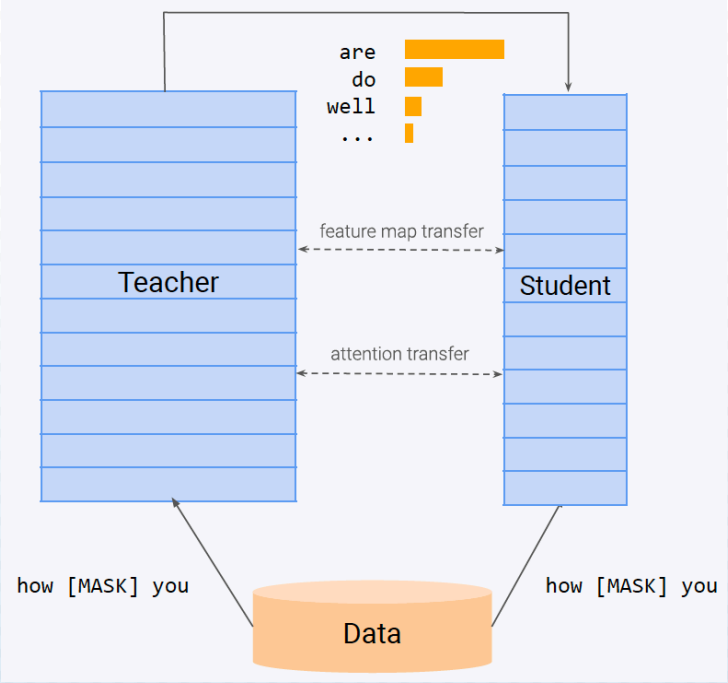
- Extreme Language Model Compression with Optimal Subwords and Shared Projections (Zhao et al., arXiv:1909.11687)
  - 减小词表（30K→5K）
  - 逐层映射（共享映射函数）
  - 最高压缩60倍

Model	Hidden Dim	Vocab Size	Compress Factor	MRPC (F1/Acc)	MNLI-m (Acc)	MNLI-mm (Acc)	SST-2 (Acc)
Teacher $BERT_{BASE}$	768	30522	1x	88.5/84.3	84.0	82.8	93.5
PKD, 6 layers (Sun et al., 2019)	768	30522	1.64x	85.0/79.9	81.5	81.0	92.0
PKD, 3 layers (Sun et al., 2019)			2.40x	80.7/72.5	76.7	76.3	87.5
NoKD Baseline	192	4928	5.74x	82.6/74.1	77.4	76.5	87.1
DualTrain				82.5/76.6	78.1	77.3	**88.4**
DualTrain + SharedProjDown				83.6/76.9	**78.2**	**77.7**	**88.4**
DualTrain + SharedProjUp				**84.9/78.5**	77.5	76.7	88.0
NoKD Baseline	96	4928	19.41x	84.6/77.3	76.2	75.1	85.4
DualTrain				**86.1/80.5**	76.1	74.7	85.4
DualTrain + SharedProjDown				83.7/77.5	**76.5**	**75.2**	**85.6**
DualTrain + SharedProjUp				84.9/78.1	76.4	**75.2**	84.7
NoKD Baseline	48	4928	61.94x	76.3/66.1	70.9	70.2	79.5
DualTrain				77.5/66.8	70.6	69.9	79.8
DualTrain + SharedProjDown				78.0/68.2	**71.3**	70.4	80.0
DualTrain + SharedProjUp				**79.3/68.6**	71.0	**70.8**	**82.2**



MobileBERT

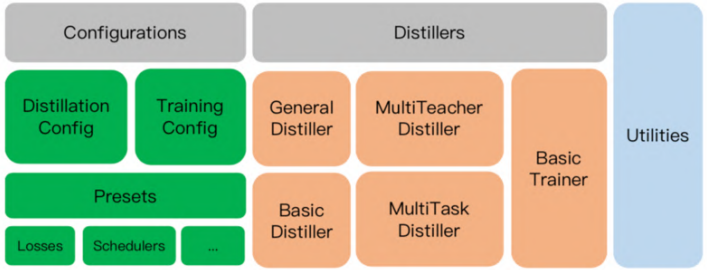
- MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices (Sun et al., arXiv:2004.02984)

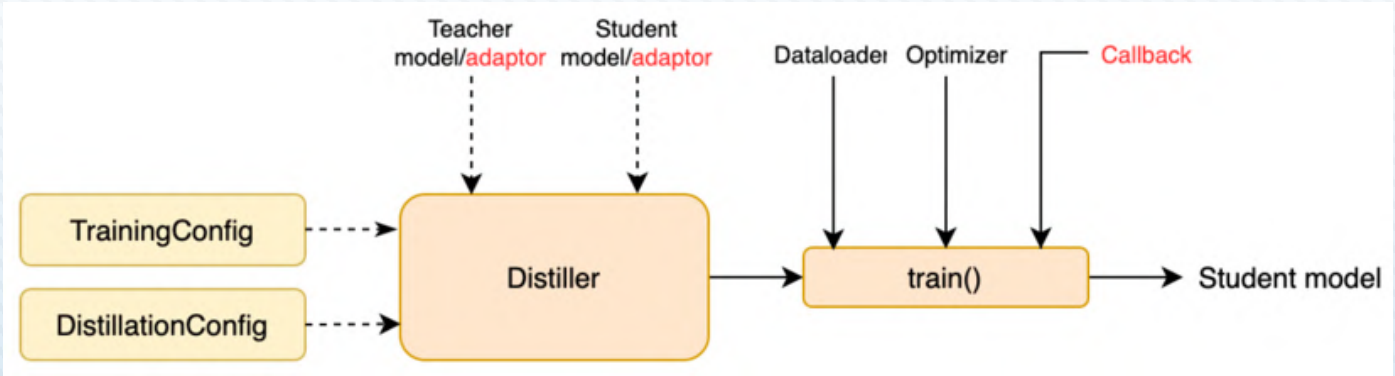


	#Params	#FLOPS	Latency	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	GLUE
				8.5k	67k	3.7k	5.7k	364k	393k	108k	2.5k	
ELMo-BiLSTM-Attn	-	-	-	33.6	90.4	84.4	72.3	63.1	74.1/74.5	79.8	58.9	70.0
OpenAI GPT	109M	-	-	47.2	93.1	87.7	84.8	70.1	80.7/80.6	87.2	69.1	76.9
$\text{BERT}_{\text{BASE}}$	109M	22.5B	342 ms	**52.1**	**93.5**	**88.9**	**85.8**	71.2	**84.6/83.4**	90.5	66.4	78.3
$\text{BERT}_{\text{BASE}}$-6L-PKD*	66.5M	11.3B	-	-	92.0	85.0	-	70.7	81.5/81.0	89.0	65.5	-
$\text{BERT}_{\text{BASE}}$-4L-PKD†*	52.2M	7.6B	-	24.8	89.4	82.6	79.8	70.2	79.9/79.3	85.1	62.3	-
$\text{BERT}_{\text{BASE}}$-3L-PKD*	45.3M	5.7B	-	-	87.5	80.7	-	68.1	76.7/76.3	84.7	58.2	-
$\text{DistilBERT}_{\text{BASE}}$-6L†	62.2M	11.3B	-	-	92.0	85.0		70.7	81.5/81.0	89.0	65.5	-
$\text{DistilBERT}_{\text{BASE}}$-4L†	52.2M	7.6B	-	32.8	91.4	82.4	76.1	68.5	78.9/78.0	85.2	54.1	-
TinyBERT*	14.5M	1.2B	-	43.3	92.6	86.4	79.9	**71.3**	82.5/81.8	87.7	62.9	75.4
$\text{MobileBERT}_{\text{TINY}}$	15.1M	3.1B	40 ms	46.7	91.7	87.9	80.1	68.9	81.5/81.6	89.5	65.1	75.8
MobileBERT	25.3M	5.7B	62 ms	50.5	92.8	88.8	84.4	70.2	83.3/82.6	90.6	66.2	77.7
MobileBERT w/o OPT	25.3M	5.7B	192 ms	51.1	92.6	88.8	84.8	70.5	84.3/**83.4**	**91.6**	**70.4**	**78.5**

TextBrewer

- TextBrewer: An Open-Source Knowledge Distillation Toolkit (Yang et al., ACL 2020[HFL])
  - 基于PyTorch的NLP知识蒸馏框架
  - 支持多种知识蒸馏方法和策略





ALBERT

- ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations (Lan et al., arXiv:1909.11942)
  - 更小的词向量维度（728 → 128）

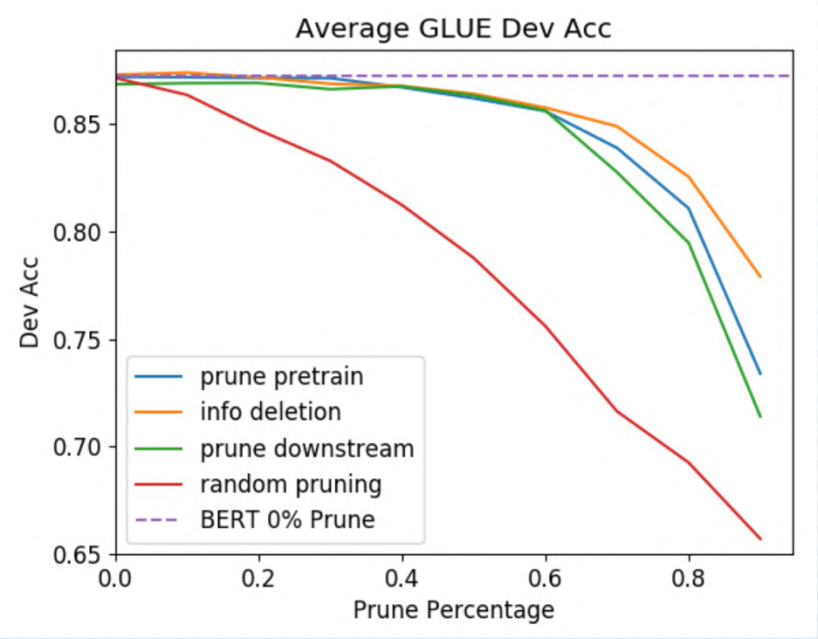
$$O(V \times H) \rightarrow O(V \times E + E \times H)$$

  - 跨层参数共享（类似循环神经网络）

  - 将下句预测（NSP）改为句子顺序预测（SOP）
    - NSP难度较低，SOP显著提升性能
  - 效果
    - 参数量大幅降低，但是不会节约时间

Compressing BERT

- Compressing BERT: Studying the effects of weight pruning on transfer learning (Gordon et al., arXiv: 2002.08307)
  - 将模型中影响较小的部分剪去
  - 探索了剪枝的时期
    - 预训练时
    - 下游任务精调时
  - 30%-40%的权重是可废弃的



Q-BERT

- Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT (Shen et al., AAAI 2020)
  - 将高精度模型用低精度来表示
  - 混合精度量化
  - 组量化

# Group	SST-2	MNLI-m/mm	CoNLL-03
Baseline	93.00	84.00/84.40	95.00
1	85.67	76.69/77.00	89.86
12	92.31	82.37/82.95	94.42
128	92.66	83.89/84.17	94.90
768 [4]	92.78	84.00/84.20	94.99

Table 2: Quantization results for $\text{BERT}_{\text{BASE}}$ on SQuAD.

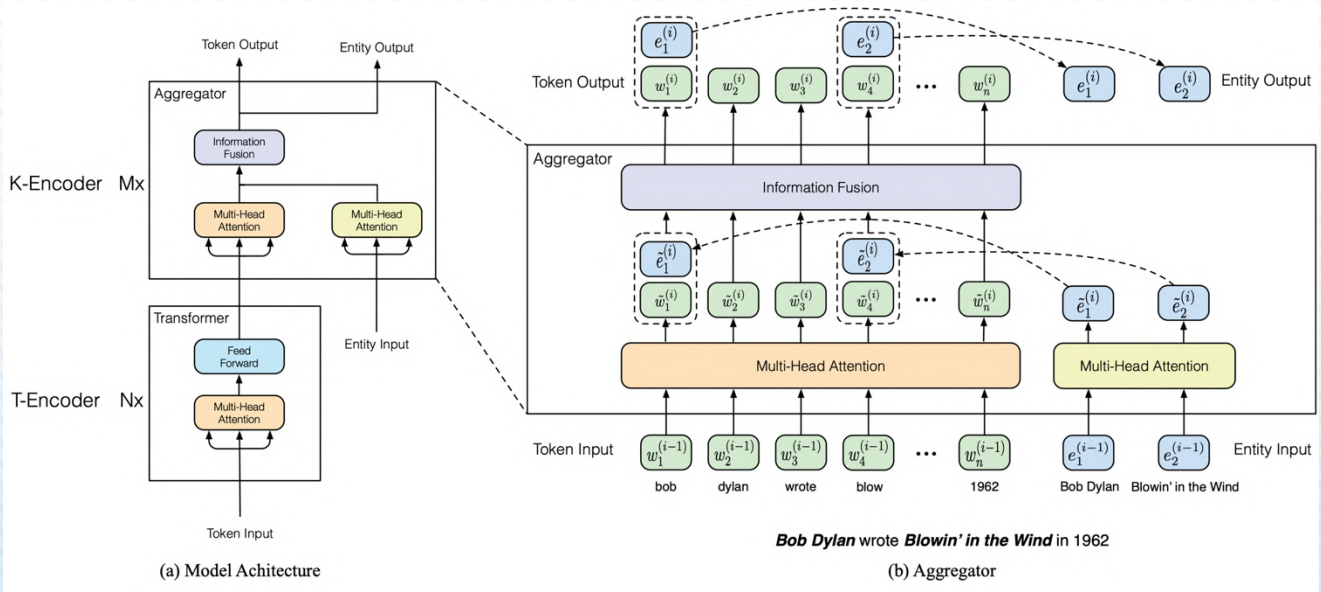
Method	w-bits	e-bits	EM	F_1	Size	Size-w/o-e
Baseline	32	32	81.54	88.69	415.4	324.5
Q-BERT	8	8	81.07	88.47	103.9	81.2
DirectQ	4	8	66.05	77.10	63.4	40.6
Q-BERT	4	8	**80.95**	**88.36**	63.4	40.6
DirectQ	3	8	46.77	59.83	53.2	30.5
Q-BERT	3	8	**79.96**	**87.66**	53.2	30.5
Q-BERT$_{\text{MP}}$	2/4 MP	8	**79.85**	**87.49**	53.2	30.5
DirectQ	2	8	4.77	10.32	43.1	20.4
Q-BERT	2	8	**69.68**	**79.60**	43.1	20.4
Q-BERT$_{\text{MP}}$	2/3 MP	8	**79.25**	**86.95**	**48.1**	**25.4**

预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

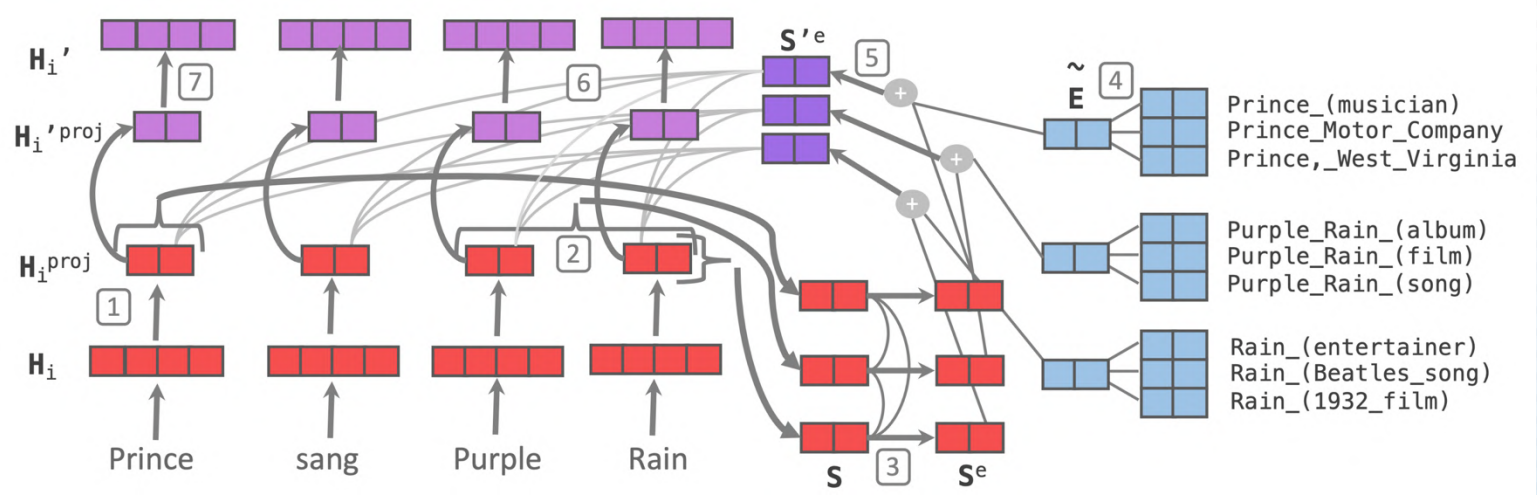
ERNIE（清华）

- ERNIE: Enhanced Language Representation with Informative Entities (Zhang et al., ACL 2019)
  - 在预训练模型中，将知识图谱中实体的表示融入文本表示



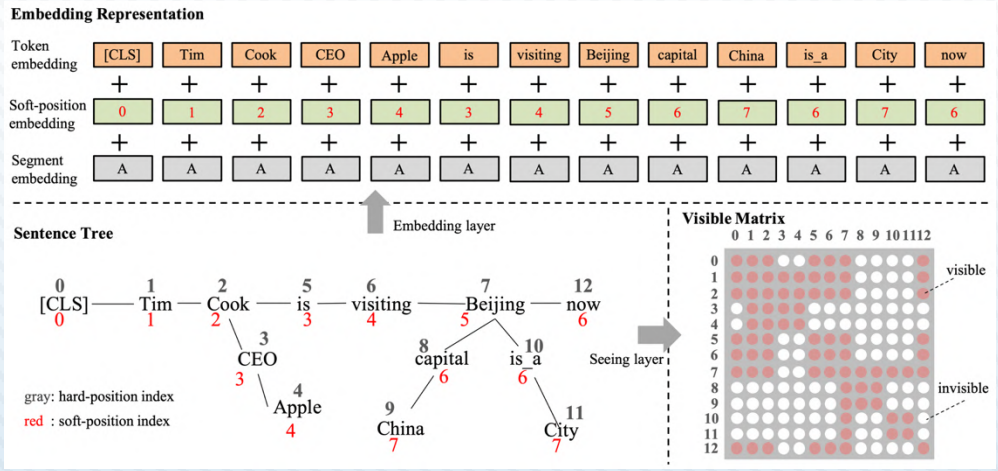
KnowBERT

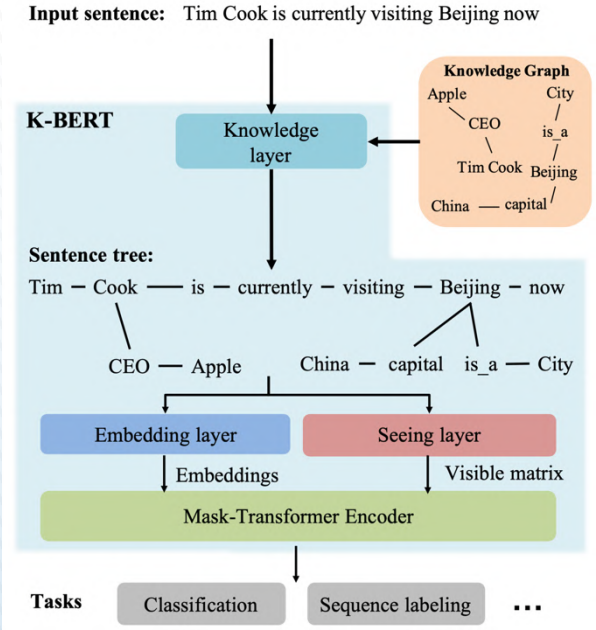
- Knowledge Enhanced Contextual Word Representations (Peters et al., EMNLP 2019)
  - 在融入知识图谱的表示时，使用注意力机制建模交互信息



K-BERT

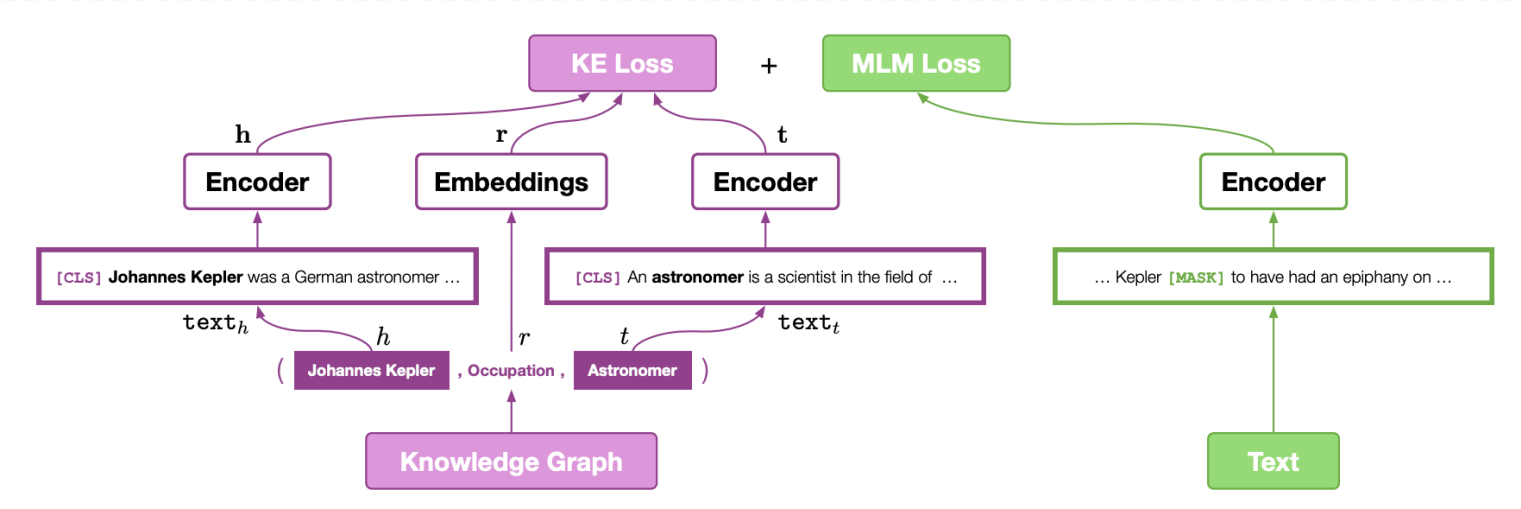
- K-BERT: Enabling Language Representation with Knowledge Graph (Liu et al., arXiv:1909.07606)
  - 在预训练模型的推理阶段引入知识图谱信息
  - 无需修改原预训练模型





KEPLER

- KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation (Wang et al., arXiv:1911.06136)
  - 知识嵌入(KE)+MLM联合训练
  - 隐式地将知识融入语言表示模型中

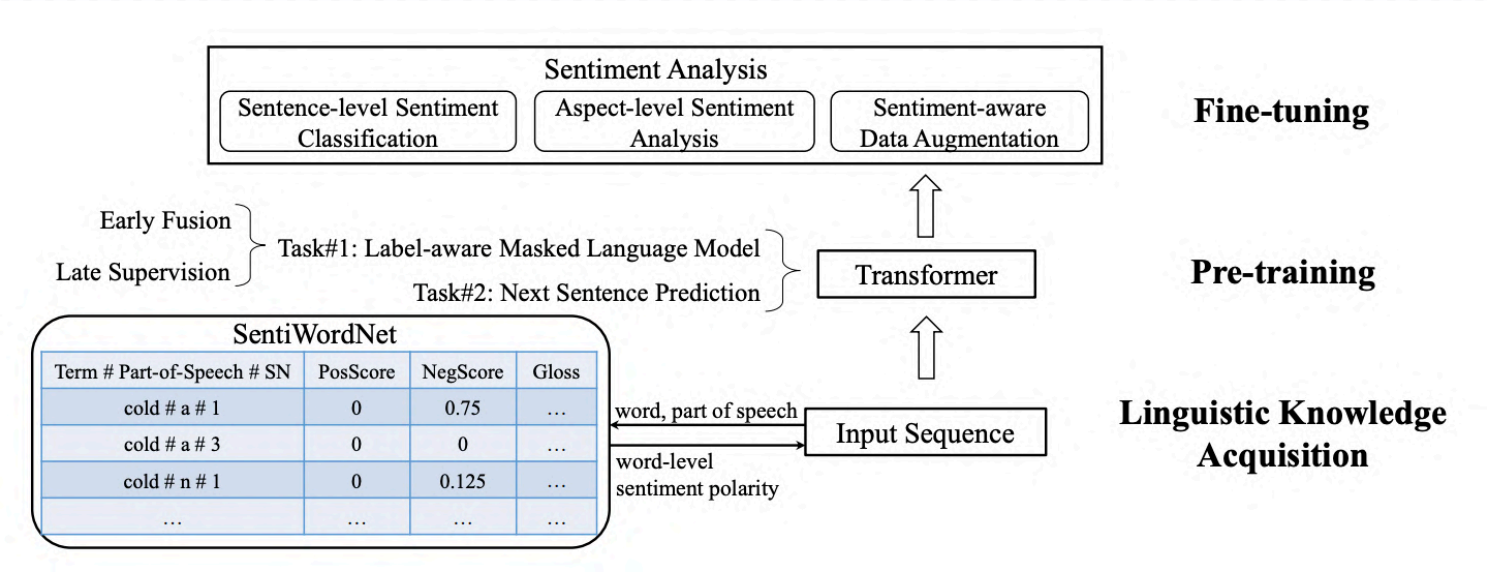


预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

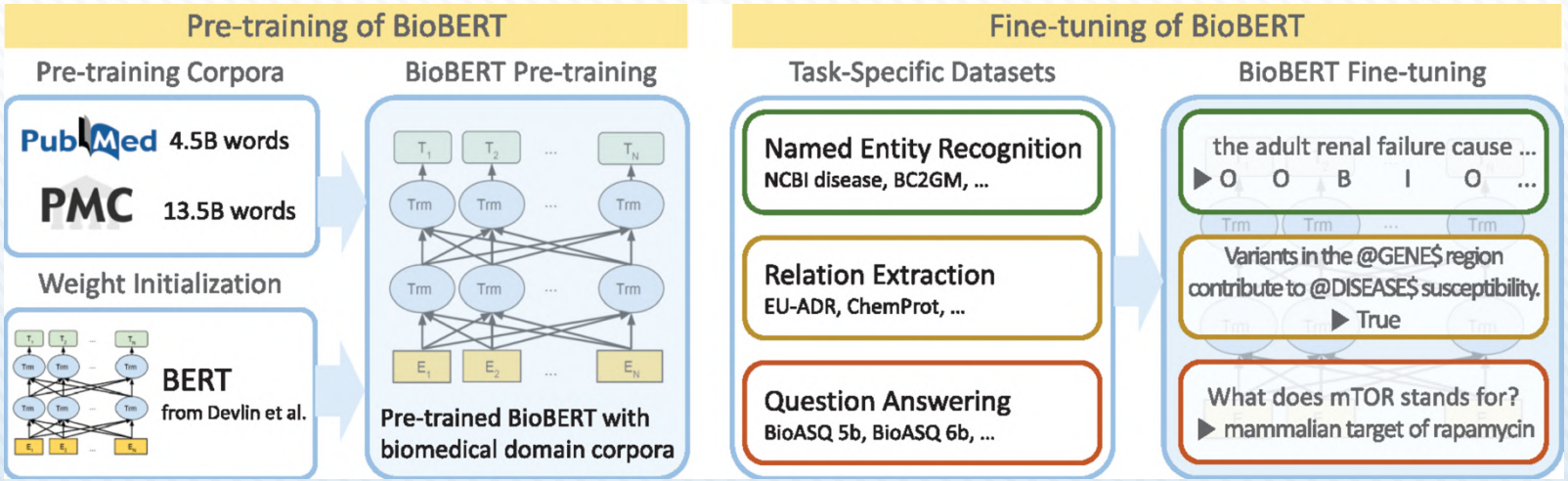
SentiLARE

- SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge (Ke et al., arXiv:1911.02493)
  - 集成了每个单词的情感极性
  - 将MLM扩展到标签感知MLM (LA-MLM)



BioBERT

- BioBERT: a pre-trained biomedical language representation model for biomedical text mining (Lee et al., Bioinformatics 2020)
  - 基于生物医学文本预训练



SciBERT

- SciBERT: A pretrained language model for scientific text (Beltagy et al., arXiv:1903.10676)
  - 基于科学文本预训练

Field	Task	Dataset	SOTA	BERT-Base		SciBERT	
				Frozen	Finetune	Frozen	Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85[7]	85.08	86.72	88.73	**90.01**
		JNLPBA (Collier and Kim, 2004)	**78.58**	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	**89.36**	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	**72.28**
	DEP	GENIA (Kim et al., 2003) - LAS	**91.92**	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	**92.84**	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	**83.64**
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	**65.71**
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	**85.42**	**85.49**
Average				73.58	77.16	76.01	79.27

Patentbert

- Patentbert: Patent classification with fine-tuning a pre-trained bert model (Lee et al., arXiv:1906.02124)
  - 用于专利分类问题的BERT预训练

	Method	Patent Data(1)	Train(2)	Test(3)	F1 (%)	Precision (%)	Recall (%)	TREC EVAL
(a)	DeepPatent	IPC+Title+Abstract	**EPO+WIPO**	EPO	N/A	**83.98**	N/A	Top 1
(b)	DeepPatent	IPC+Title+Abstract	**2006~2014**	2015-A	N/A	**73.88**	N/A	Top 1
(c)	DeepPatent	IPC+Title+Abstract	**EPO+WIPO**	EPO	**55.09**	45.79	75.46	**Top 4**
(d)	**DeepPatent**	IPC+Title+Abstract	**2006~2014**	2015-A	**< 45**	< 35	< 74	Top 5
(e)	**PatentBERT**	IPC+Title+Abstract	2006~2014	2015-A	**46.85**	32.19	86.06	Top 5
(f)	PatentBERT	IPC+Title+Abstract	2006~2014	2015-A	**64.91**	80.61	54.33	**Top 1**
(g)	PatentBERT	IPC+**Claim**	2006~2014	2015-A	**63.74**	79.14	53.36	Top 1
(h)	PatentBERT	**CPC**+Claim	2006~2014	2015-A	**66.83**	**84.26**	**55.38**	**Top 1**
(i)	PatentBERT	CPC+Claim	2006~2014	**2015-B**	**66.80**	84.24	55.35	Top 1
(j)	PatentBERT	CPC+Claim	**2000~2014**	2015-B	**66.71**	84.95	54.92	Top 1
(k)	PatentBERT	CPC+Claim	2000~2014	**2016**	65.89	84.89	53.84	Top 1
(l)	PatentBERT	CPC+Claim	2000~2014	**2017**	65.35	83.97	53.49	Top 1

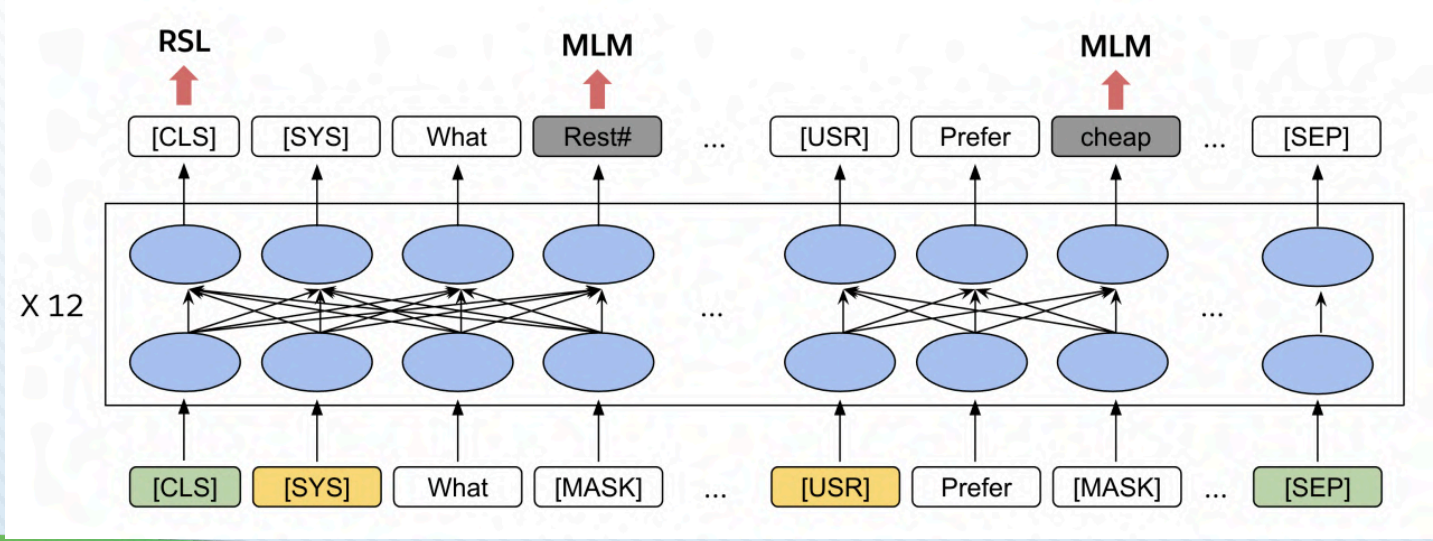
(1) IPC subclass level: 632 labels. CPC subclass level: 656 labels

(2) Training dataset size:
- EPO: 580,546 patents. WIPO: 161,551 patents.
- USPTO-2M: 2,000,147 patents by the DeepPatent (2006~2015, from USPTO)
- USPTO-3M: 3,050,615 patents, our new dataset with SQL statements (2000~2015, from Google Patents Public Datasets on BigQuery)
- 2006~2014: 1,950,247 patents out of USPTO-2M for DeepPatent. 1,933,105 patents for PatentBERT. Minor discrepancy exists due to different data sources and probably preprocessing criteria.
- 2000~2014 : 2,900,615 patents out of USPTO-3M for PatentBERT

(3) Testing dataset:
- EPO: 1,350 patents
- 2015-A: 49,900 patents out of USPTO-2M for DeepPatent. 49,670 patents for PatentBERT (out of USPTO-3M and based on DeepPatent's list of test patents)
- 2015-B: 150,000 of the 298,559 patents in 2015 (from USPTO-3M)
- 2016: 150,000 of the 298,559 patents in 2016 (from BigQuery)
- 2017: 150,000 of the 298,559 patents in 2017-01~2017-08 (from BigQuery)

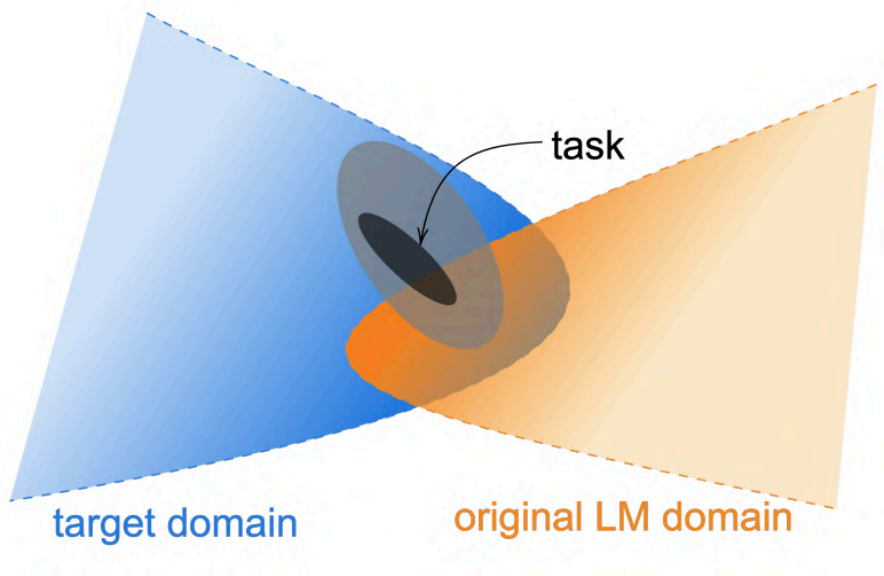
ToD-BERT

- ToD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogues (Wu et al., arXiv:2004.06871)
  - 选用9个任务型对话领域数据集进行预训练
  - MLM+Response Selection多任务训练



Don't Stop Pretraining

- Don't Stop Pretraining: Adapt Language Models to Domains and Tasks (Suchin et al., ACL 2020)
  - 在预训练模型的基础上，增加大量领域内未标注文本继续训练语言模型
  - 再在指定任务的数据集上进行微调

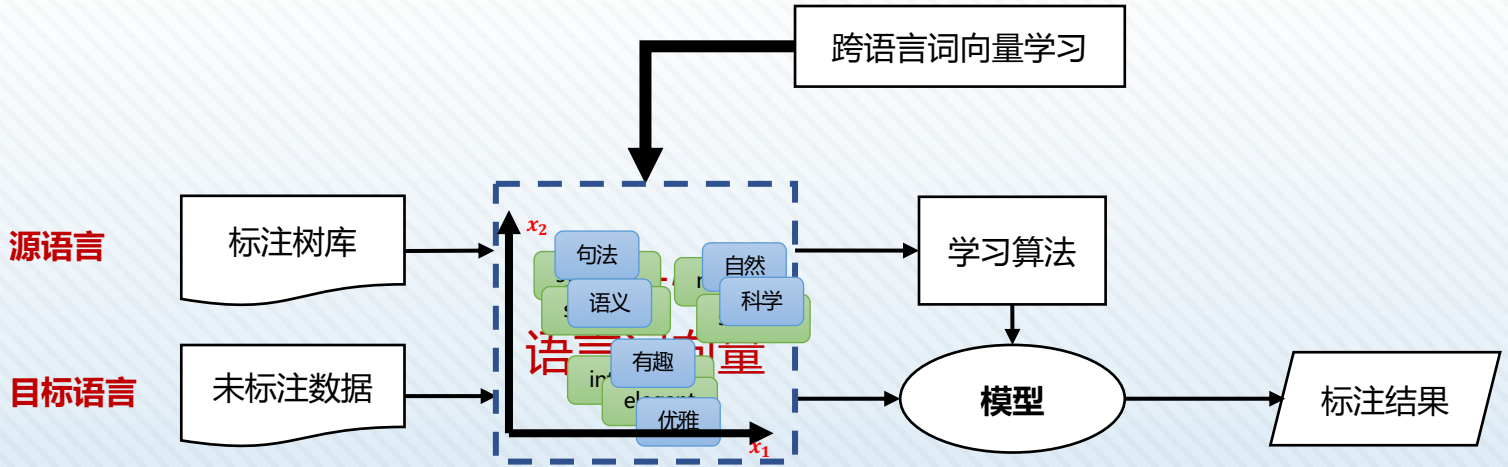


预训练模型的进展

- 更多预训练任务
  - 文本判别任务
  - 文本生成任务
- 预训练模型调优
  - 更精细的调参
  - 新的模型结构
  - 模型压缩与加速
- 融入知识图谱
- 特定领域预训练
- 跨语言与跨模态

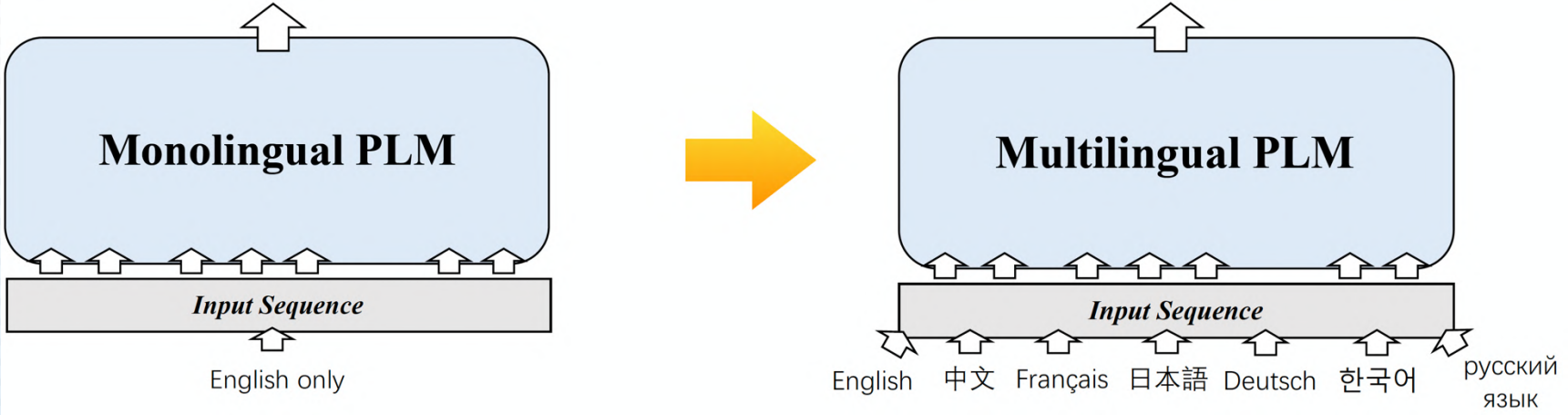
传统跨语言方法

- 以跨语言句法分析为例
  - Cross-Lingual Dependency Parsing Based on Distributed Representations (Guo et al., ACL 2015[SCIR])
    - 基于“静态”词向量



多语言预训练语言模型

- 基于“动态”词向量
- 统一的多语言预训练语言模型
- 实现“跨语言”能力

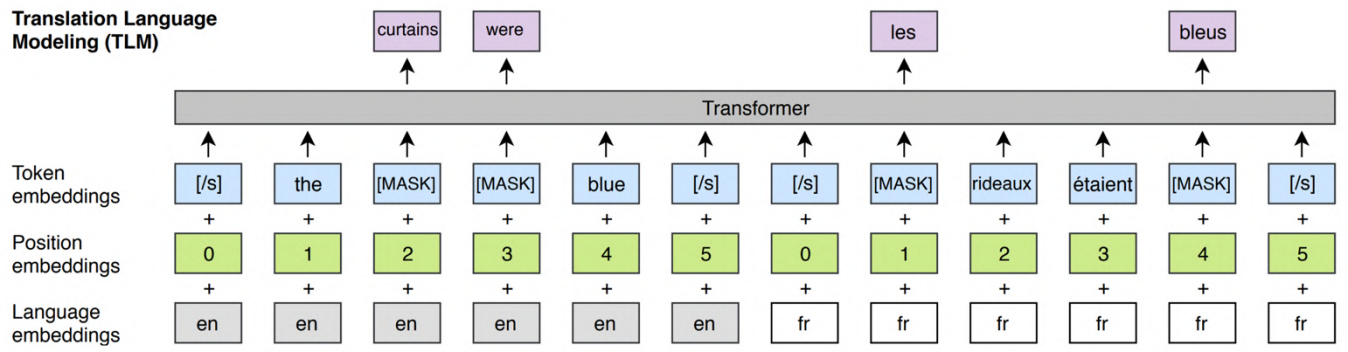


多语言BERT

- Multilingual BERT (M-BERT) (Devlin et al., NAACL 2019)
  - Google官方发布的104种语言BERT
  - 直接使用104种语言的Wikipedia单语数据训练
    - 语言之间共享相同的Word-Piece
    - 很多语言混杂在一起（Code-switching）
  - 在多个跨语言任务上表现优异
  - 问题
    - 不适用距离较远的语言对
    - 准确率不如单语BERT

跨语言预训练语言模型

- XLM: Cross-lingual Language Model (Lample and Conneau, arXiv:1901.07291)
  - 将互为翻译的句子作为BERT结构的输入
  - 随机Mask句对中的双语词
  - 问题
    - 依赖大规模双语语料库
    - 需要大规模计算资源



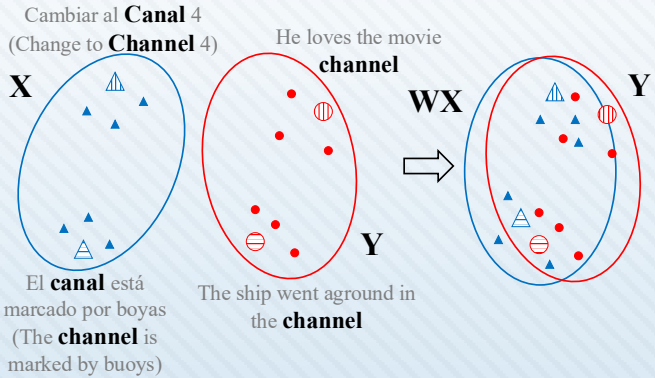
XLM-R (XLM with RoBERTa)

- Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., ACL 2020)
  - 只使用单语数据，取消了平行数据的限制
  - 更大的模型（RoBERTa）
  - 更多的数据（尤其是对小语种）

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{Base}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**

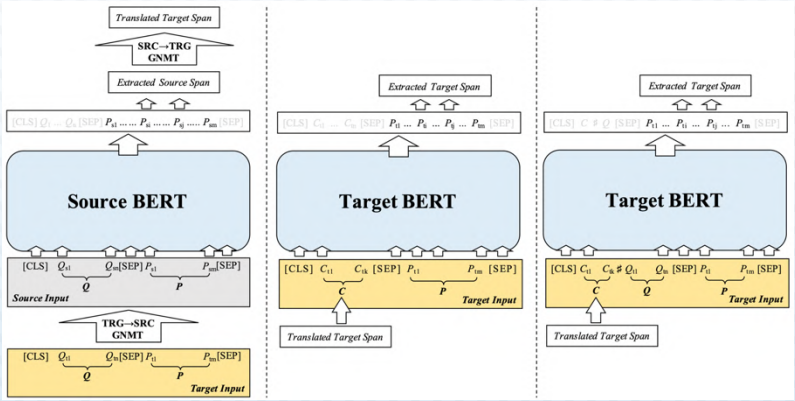
跨语言映射BERT

- Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing (Wang et al., EMNLP 2019 SCIR)
  - 直接使用单语言预训练的BERT
  - 假设双语句对中互为翻译的词具有相同的词向量
  - 通过线性变换，将目标语言的上下文词向量映射到源语言
  - 优势
    - 仅需少量双语语料库和计算资源

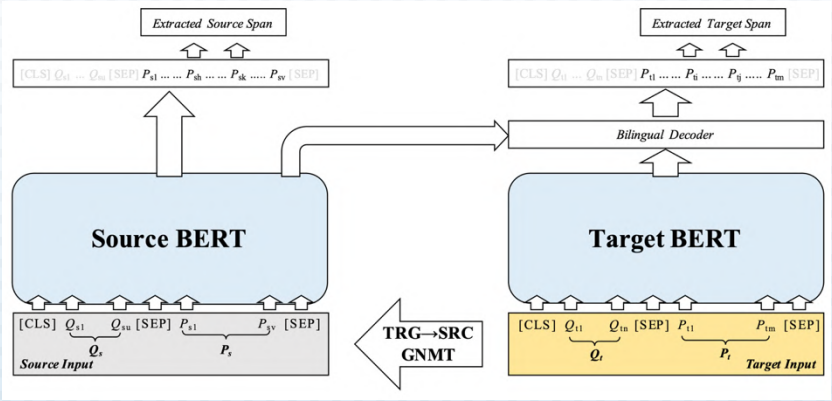


跨语言阅读理解

- Cross-Lingual MRC (Cui et al., EMNLP 2019HFL)
  - 除英语外其它语言缺乏大规模阅读理解数据
  - 将英语阅读理解模型应用于其它语言
  - 方法
    - 改进回翻技术
    - Dual BERT



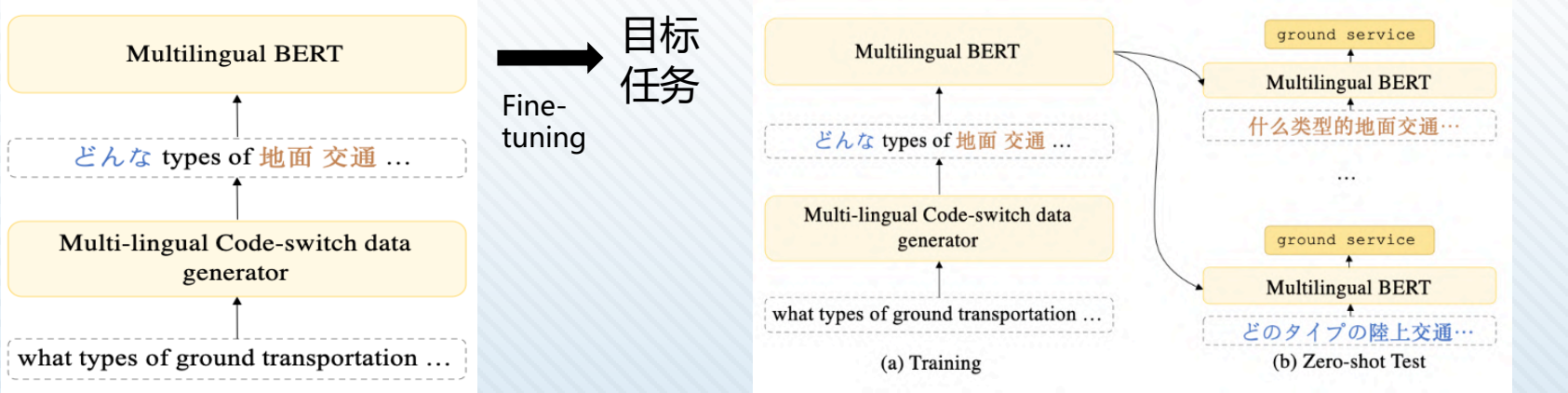
改进回翻技术



Dual BERT

Multi-Lingual Code-Switching

- CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP (Qin et al., IJCAI 2020[SCIR])
  - 一种数据增强框架，生成多语言code-switch训练数据
  - 不依赖于双语句对进行训练，一次训练能运用到多个目标语言



模型训练　　模型测试

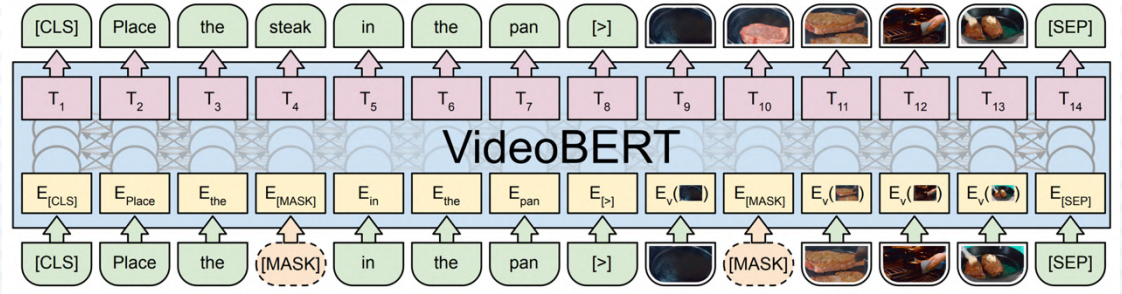
跨语言预训练模型Leaderboard

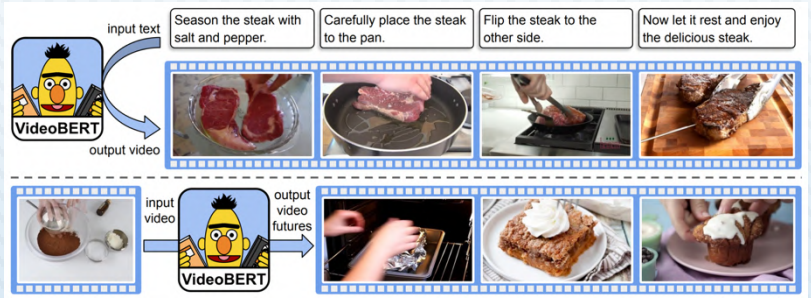
- XTREME
  - **(X)** Cross-Lingual **Tr**ansfer **E**valuation of **M**ultilingual **E**ncoders
  - https://sites.research.google/xtreme
  - 9项自然语言处理任务
  - 12个语族的40种语言

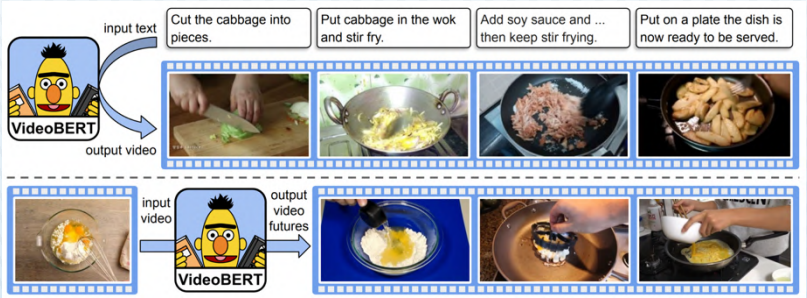
Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
2	Polyglot	MLNLC	ByteDance	Nov 13, 2020	77.8	87.8	72.9	67.4	88.3
3	VECO	DAMO NLP Team	Alibaba	Sep 29, 2020	77.2	87.0	70.4	68.0	88.1
4	FILTER	Dynamics 365 AI Research	Microsoft	Sep 8, 2020	77.0	87.5	71.9	68.5	84.4
5	X-STILTs	Phang et al.	New York University	Jun 17, 2020	73.5	83.9	69.4	67.2	76.5
6	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
7	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7
8	MMTE	XTREME Team	Alphabet, CMU	-	59.3	74.3	65.3	52.3	48.9
9	RemBERT	Anonymous2	Anonymous2	-	56.1	84.1	73.3	68.6	-

跨模态BERT

- VideoBERT: A Joint Model for Video and Language Representation Learning (Sun et al., ICCV 2019)
  - 类似XLM，将文本和视频对作为BERT的输入，同时Mask词以及图像块

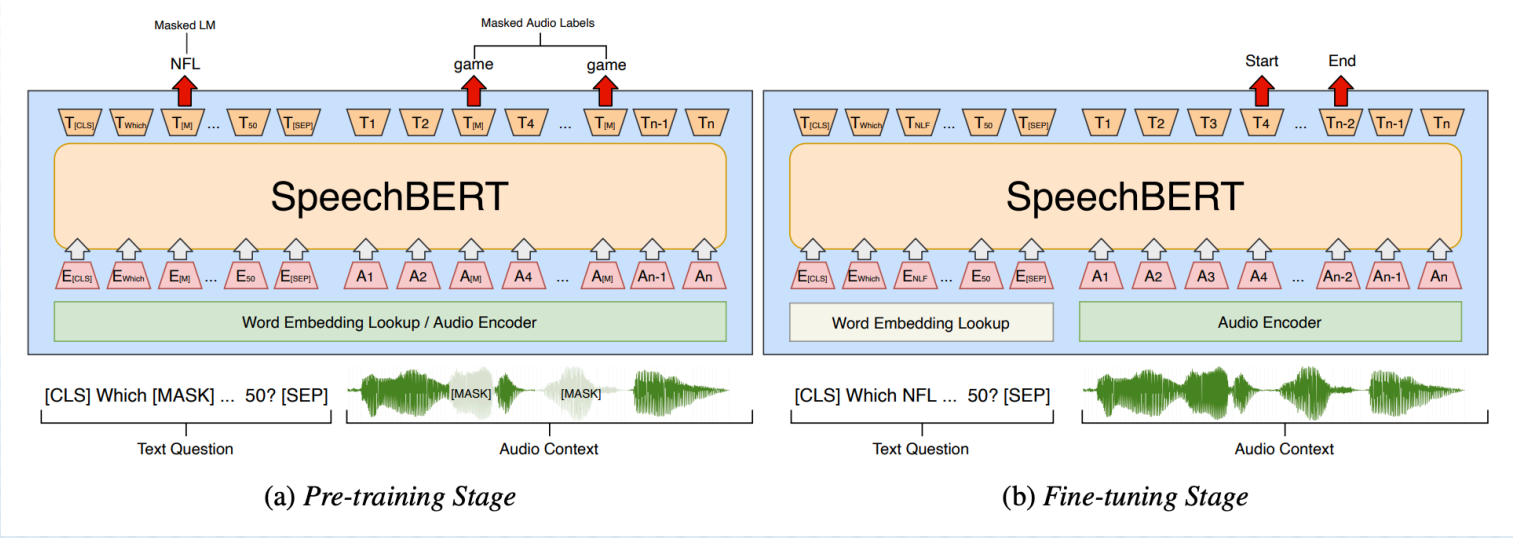






跨模态BERT

- SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering (Chuang et al., arXiv:1910.11559)
  - 将文本和音频对作为BERT的输入，同时Mask词以及音频片段
  - 在跨模态QA上精调



各种跨模态BERT对比

- VL-BERT: Pre-training of Generic Visual-Linguistic Representations (Su et al., arXiv:1908.08530)

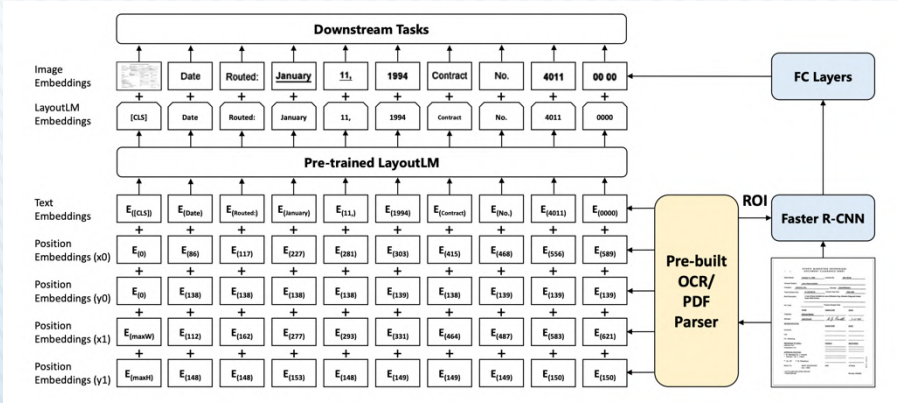
	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Hao Tan, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
Works in Progress	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

LayoutLM

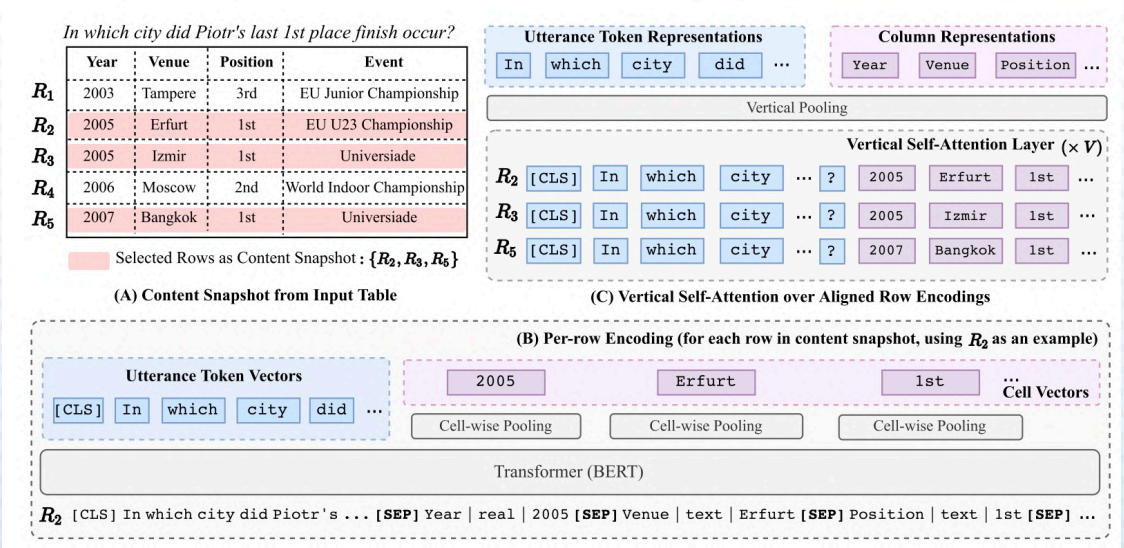
- LayoutLM: Pre-training of Text and Layout for Document Image Understanding (Xu et al., KDD 2020)
  - 为了结合文档结构和视觉信息，在现有的预训练模型基础上添加以下信息
    - OCR获得文本Bounding Box → 2-D Position Embedding
    - 文本对应候选框 → Image Embedding
  - 自监督预训练任务
    - 遮罩式视觉语言模型（学习模态对齐关系）
    - 多标签文档分类 （捕捉文档类型信息）

Modality	Model	Precision	Recall	F1
Text only	$\text{BERT}_{\text{BASE}}$	0.5469	0.671	0.6026
	$\text{RoBERTa}_{\text{BASE}}$	0.6349	0.6975	0.6648
	$\text{BERT}_{\text{LARGE}}$	0.6113	0.7085	0.6563
	$\text{RoBERTa}_{\text{LARGE}}$	0.678	0.7391	0.7072
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.665	0.7355	0.6985
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.7377	0.782	0.7592
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.7597	0.8155	0.7866
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7076	0.7695	0.7372
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.7194	0.7780	0.7475
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.7171	0.805	0.7585
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	0.7536	0.806	0.7789
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7101	0.7815	0.7441
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**0.7677**	**0.8195**	**0.7927**



TaBERT

- TABERT: Pretraining for Joint Understanding of Textual and Tabular Data (Yin, et al., ACL 2020)
  - 对结构化表格信息和文本信息进行建模，让模型在预训练阶段进行多模态对齐
  - 达到WikiTableQuestion数据集的SOTA，从44.5提高到51.8



Previous Systems on WikiTableQuestions

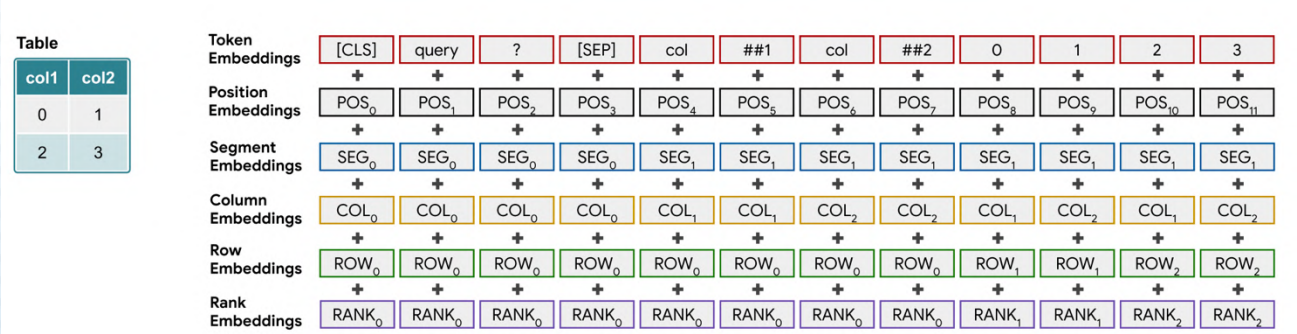
Model	DEV	TEST
Pasupat and Liang (2015)	37.0	37.1
Neelakantan et al. (2016)	34.1	34.2
Ensemble 15 Models	37.5	37.7
Zhang et al. (2017)	40.6	43.7
Dasigi et al. (2019)	43.1	44.3
Agarwal et al. (2019)	43.2	44.1
Ensemble 10 Models	–	46.9
Wang et al. (2019b)	43.7	44.5

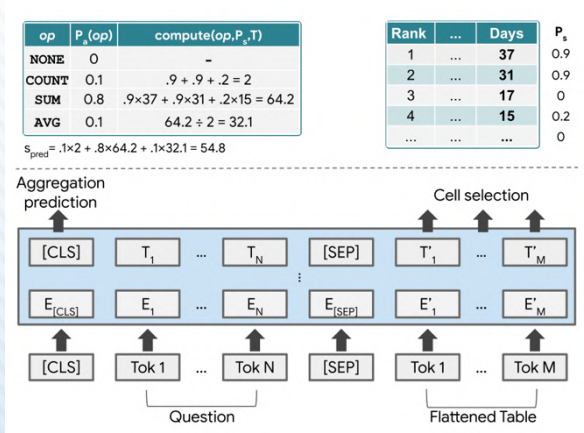
Our System based on MAPO (Liang et al., 2018)

	DEV	Best	TEST	Best
Base Parser†	42.3 ±0.3	42.7	43.1 ±0.5	43.8
w/ $\text{BERT}_{\text{Base}}$ (K = 1)	49.6 ±0.5	50.4	49.4 ±0.5	49.2
– content snapshot	49.1 ±0.6	50.0	48.8 ±0.9	50.2
w/ $\text{TABERT}_{\text{Base}}$ (K = 1)	51.2 ±0.5	51.6	50.4 ±0.5	51.2
– content snapshot	49.9 ±0.4	50.3	49.4 ±0.4	50.0
w/ $\text{TABERT}_{\text{Base}}$ (K = 3)	51.6 ±0.5	52.4	51.4 ±0.3	51.3
w/ $\text{BERT}_{\text{Large}}$ (K = 1)	50.3 ±0.4	50.8	49.6 ±0.5	50.1
w/ $\text{TABERT}_{\text{Large}}$ (K = 1)	51.6 ±1.1	52.7	51.2 ±0.9	51.5
w/ $\text{TABERT}_{\text{Large}}$ (K = 3)	**52.2** ±0.7	**53.0**	**51.8** ±0.6	**52.3**

TAPAS

- TAPAS: Weakly Supervised Table Parsing via Pre-training (Herzig et al., ACL 2020)
  - 预训练阶段，利用位置向量融入了表格的结构化信息
  - 微调阶段，设计了以下预测函数，充分利用任务提供的弱监督信号
    - 单元格选择
    - 操作符预测





大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

中文BERT-wwm

- Pre-Training with Whole Word Masking for Chinese BERT (Yang et al., arXiv:1906.08101 HFL)
  - https://github.com/ymcui/Chinese-BERT-wwm
  - 全词遮盖
  - 使用LTP分词

[Original Sentence] 使用语言模型来预测下一个词的probability。 **[Original Sentence with CWS]** 使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。
[Original BERT Input] 使 用 语 言 **[MASK]** **型** 来 **[MASK]** **测** 下 一 个 词 的 **pro** **[MASK]** **##lity** 。 **[Whold Word Masking Input]** 使 用 语 言 **[MASK]** **[MASK]** 来 **[MASK]** **[MASK]** 下 一 个 词 的 **[MASK]** **[MASK]** **[MASK]** 。

更多中文预训练模型

- ERNIE: Enhanced Representation through kNowledge IntEgration
  - https://github.com/PaddlePaddle/ERNIE
- NEZHA: NEural ContextualiZed Representation for CHinese LAnguage Understanding
  - https://github.com/huawei-noah/Pretrained-Language-Model
- ZEN: Pre-training Chinese (Z) Text Encoder Enhanced by N-gram Representations
  - https://github.com/sinovation/ZEN

MacBERT

- MacBERT: MLM as correction BERT (Cui et al., Findings of EMNLP 2020[HFL])
  - https://github.com/ymcui/MacBERT
  - 对多种中文预训练模型进行了详细的比较
  - 使用近义词代替[Mask]符号
  - 解决预训练与精调阶段输入不一致问题

	BERT	ERNIE	XLNet	RoBERTa	ALBERT	ELECTRA	MacBERT
Type	AE	AE	AR	AE	AE	AE	AE
Embeddings	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P
Masking	T	T/E/Ph	-	T	T	T	WWM/NM
LM Task	MLM	MLM	PLM	MLM	MLM	Gen-Dis	Mac
Paired Task	NSP	NSP	-	-	SOP	-	SOP

MacBERT

- 哈工大讯飞联合实验室（HFL）在GLUE基准测试集中**排名第一**
  - MacALBERT: Mac + ALBERT-xxlarge
  - DKM: Dynamic keyword matching

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
2	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	90.7	97.4	91.2	94.5	49.1
3	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
4	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
5	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1

大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

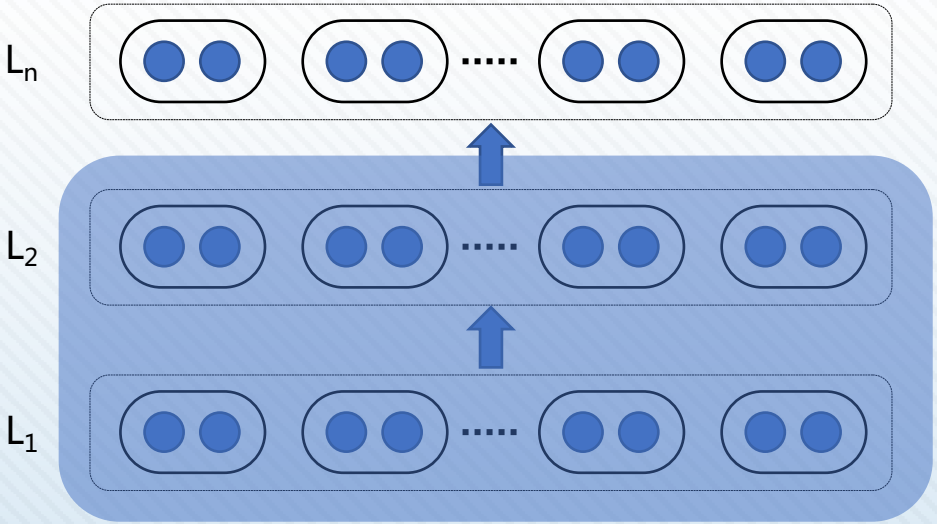
是否需要精调（Fine-tune）？

- To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks (Peters et al., arXiv:1903.05987)
  - 如果不进行Fine-tune ❄️，则需要任务相关的复杂模型
  - 如果进行Fine-tune 🔥，则任务相关模型要尽量简单

Pretraining	Adaptation	NER CoNLL 2003	SA SST-2	Nat. lang. inference MNLI	Nat. lang. inference SICK-E	Semantic textual similarity SICK-R	Semantic textual similarity MRPC	Semantic textual similarity STS-B
Skip-thoughts	❄️	-	81.8	62.9	-	86.6	75.8	71.8
ELMo	❄️	91.7	**91.8**	**79.6**	**86.3**	**86.1**	**76.0**	**75.9**
	🔥	**91.9**	91.2	76.4	83.3	83.3	74.7	75.5
	Δ=🔥-❄️	0.2	-0.6	-3.2	-3.3	-2.8	-1.3	-0.4
BERT-base	❄️	92.2	93.0	**84.6**	84.8	86.4	78.1	82.9
	🔥	**92.4**	**93.5**	**84.6**	**85.8**	**88.7**	**84.8**	**87.1**
	Δ=🔥-❄️	0.2	0.5	0.0	1.0	2.3	6.7	4.2

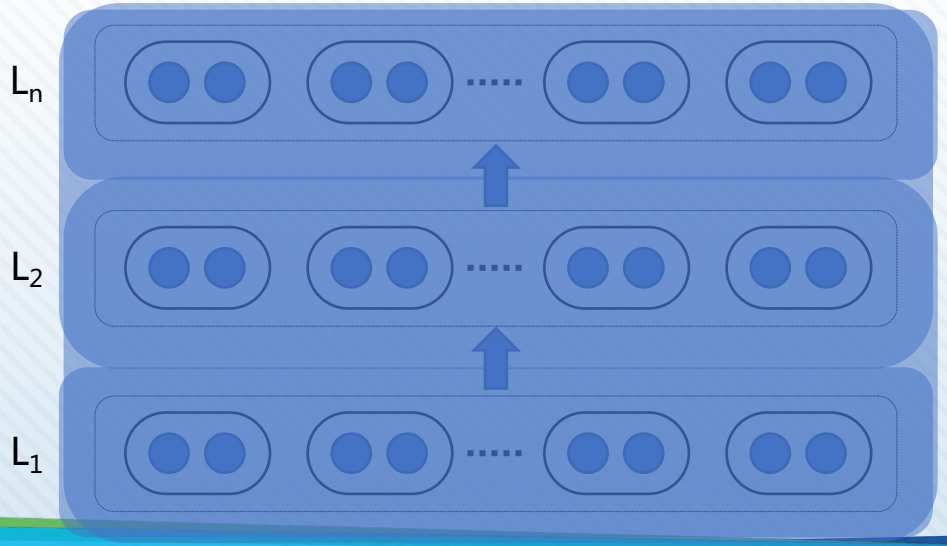
更多精调方法

- 目标：既要适应目标任务，又要避免重写预训练模型
- 方法
  - 只精调最后一层，固定其它层 (Long et al., ICML 2015)



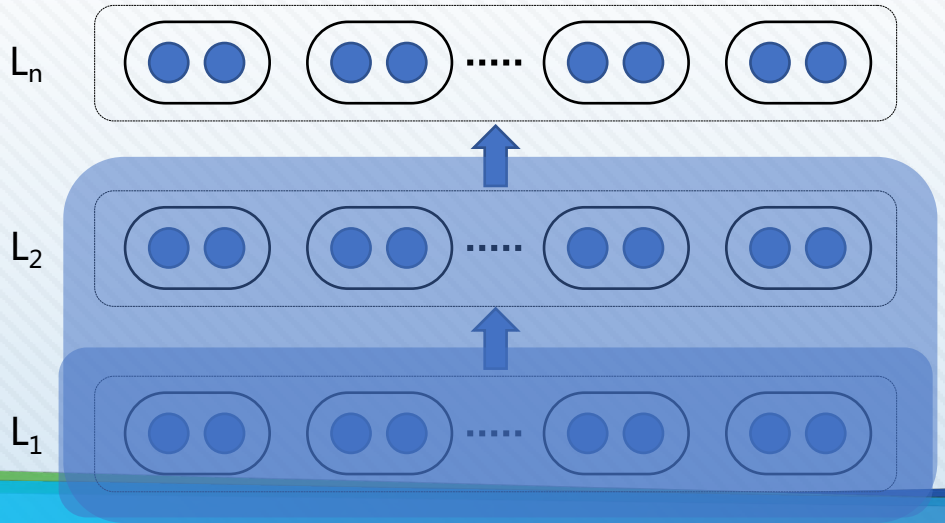
更多精调方法

- 目标：既要适应目标任务，又要避免重写预训练模型
- 方法
  - 只精调最后一层，固定其它层 (Long et al., ICML 2015)
  - 每次只精调一层，固定其它层 (Felbo et al., EMNLP 2017)



更多精调方法

- 目标：既要适应目标任务，又要避免重写预训练模型
- 方法
  - 只精调最后一层，固定其它层 (Long et al., ICML 2015)
  - 每次只精调一层，固定其它层 (Felbo et al., EMNLP 2017)
  - 自顶向下逐层解冻 (Howard and Ruder, ACL 2018)



更多精调方法

- 目标：既要适应目标任务，又要避免重写预训练模型
- 方法
  - 只精调最后一层，固定其它层 (Long et al., ICML 2015)
  - 每次只精调一层，固定其它层 (Felbo et al., EMNLP 2017)
  - 自顶向下逐层解冻 (Howard and Ruder, ACL 2018)
    - 其它策略
      - 学习率预热
      - 二次预训练：在目标领域未标注数据上精调语言模型
  - 将目标模型每层的参数和激活与预训练模型进行比较，作为额外损失 (Wiese et al., CoNLL 2017)

更多精调方法

- Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting (Chen et al., EMNLP 2020[SCIR])
  - 预训练+精调模式会面临对于预训练知识灾难性遗忘的问题
  - 提出且回忆且学习：预训练模拟机制
    - 精调模型与预训练模型的参数相似度作为正则化项
  - 对Adam优化器进行简单的改造
    - https://github.com/Sanyuan-Chen/RecAdam
  - 在8个不同NLP任务上取得了稳定提升

Algorithm 1 Adam and RecAdam

given initial learning rate $\alpha \in \mathbb{R}$, momentum factors $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, pretrained parameter vector $\theta^* \in \mathbb{R}^n$, coefficient of quadratic penalty $\gamma \in \mathbb{R}$, annealing coefficient in objective function $\lambda(t) = 1/(1 + \exp(-k \cdot (t - t_0)), k \in \mathbb{R}, t_0 \in \mathbb{N}$

initialize timestep $t \leftarrow 0$, parameter vector $\theta_{t=0} \in \mathbb{R}^n$, first moment vector $m_{t=0} \leftarrow 0$, second moment vector $v_{t=0} \leftarrow 0$, schedule multiplier $\eta_{t=0} \in \mathbb{R}$

repeat

  $t \leftarrow t + 1$

  $\nabla f_t(\theta_{t-1}) \leftarrow \text{SelectBatch}(\theta_{t-1})$ ▷ select batch and return the corresponding gradient

  $g_t \leftarrow \lambda(t) \nabla f_t(\theta_{t-1}) + (1 - \lambda(t))\gamma(\theta_{t-1} - \theta^*)$

  $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$ ▷ here and below all operations are element-wise

  $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

  $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ ▷ β_1 is taken to the power of t

  $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ ▷ β_2 is taken to the power of t

  $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$ ▷ can be fixed, decay, or also be used for warm restarts

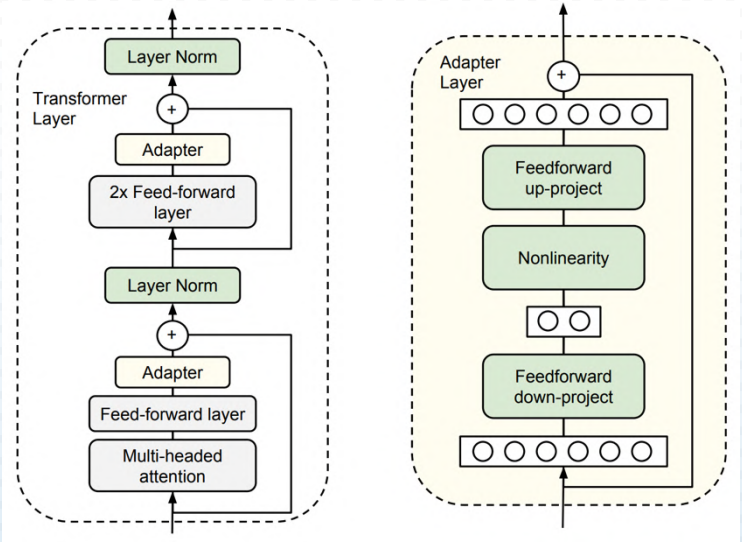
  $\theta_t \leftarrow \theta_{t-1} - \eta_t \left(\lambda(t) \alpha \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) + (1 - \lambda(t))\gamma(\theta_{t-1} - \theta^*) \right)$

until *stopping criterion is met*

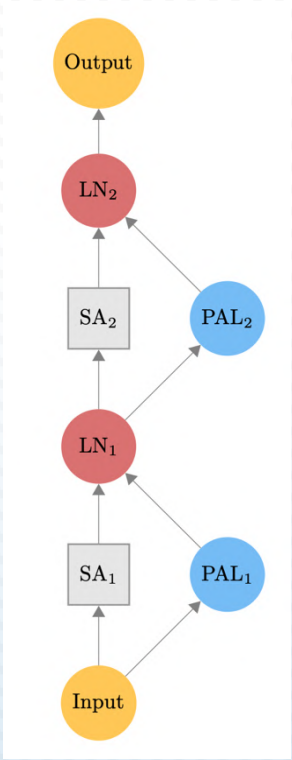
return optimized parameters θ_t

更多精调方法

- 在Transformer中增加适配器（Adapter）



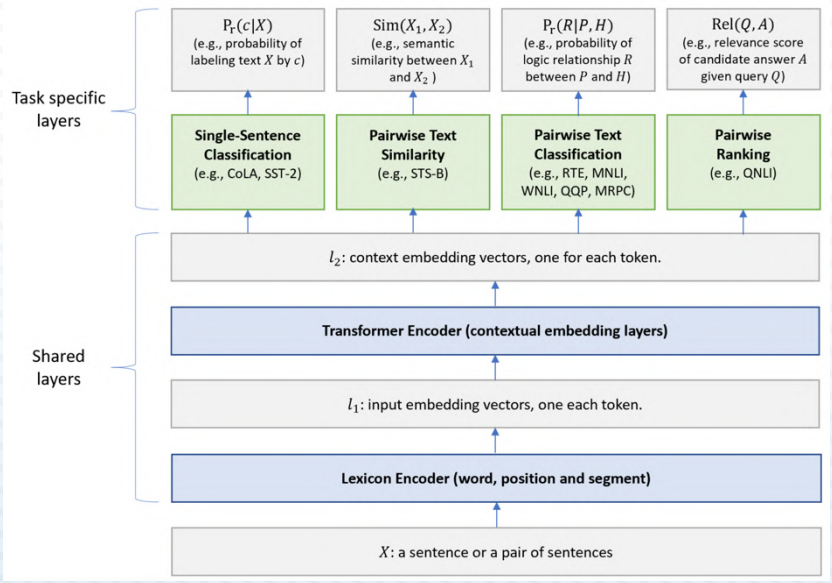
(Houlsby et al., ICML 2019)



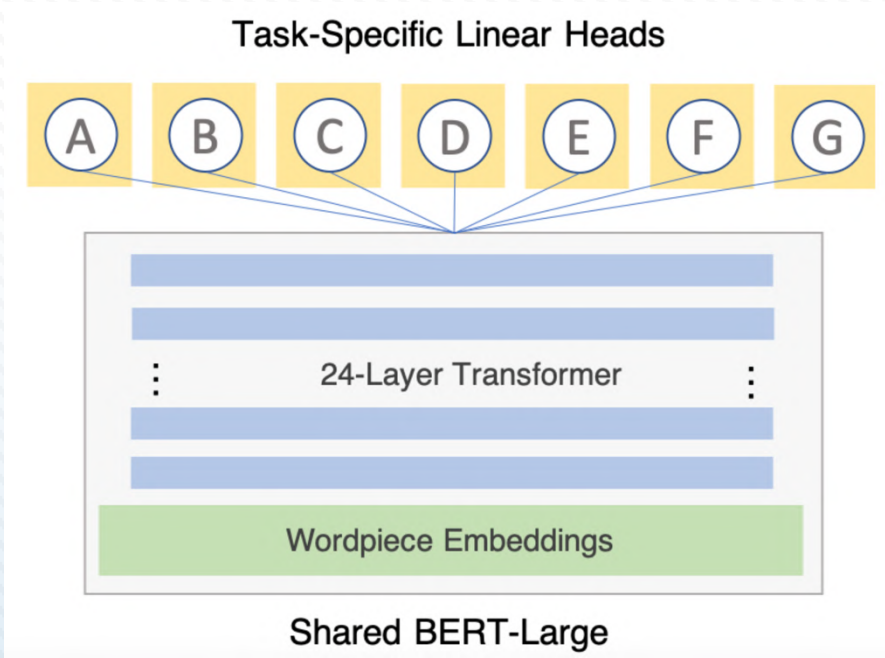
(Stickland and Murray, ICML 2019)

多任务学习

- 使用多任务学习框架，综合利用多种类型数据



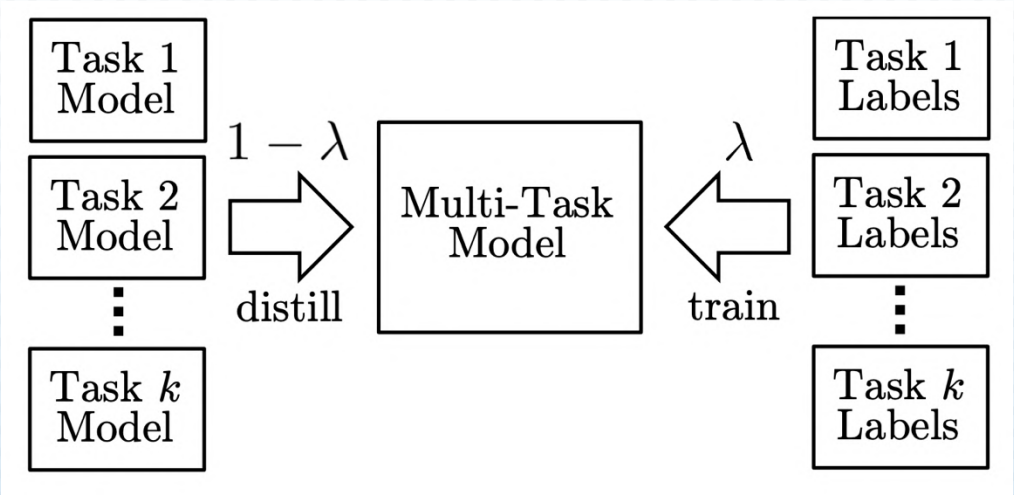
(Liu et al., ACL 2019)



https://dawn.cs.stanford.edu/2019/03/22/glue/

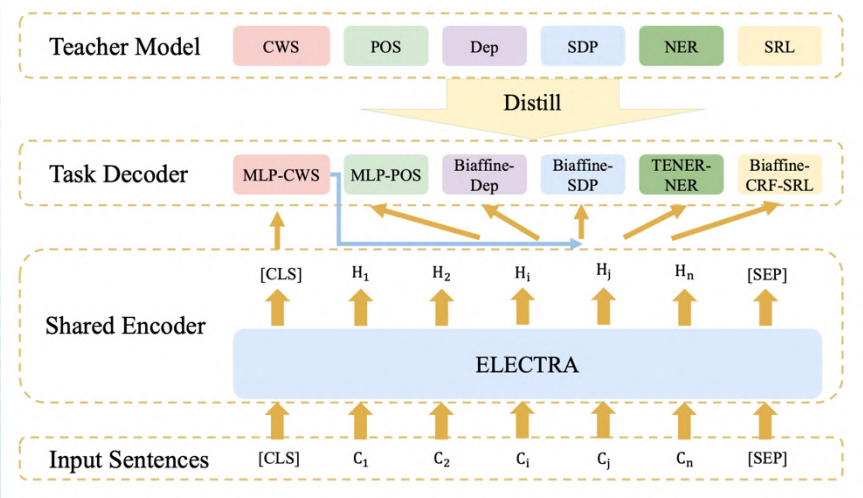
多任务学习

- BAM! Born-Again Multi-Task Networks for Natural Language Understanding (Clark et al., ACL 2019)
  - 多任务学习往往较难同时提高全部任务的性能
  - 采用知识蒸馏的技术，MTL模型学习单模型的输出概率
  - 同时提高多项任务的性能



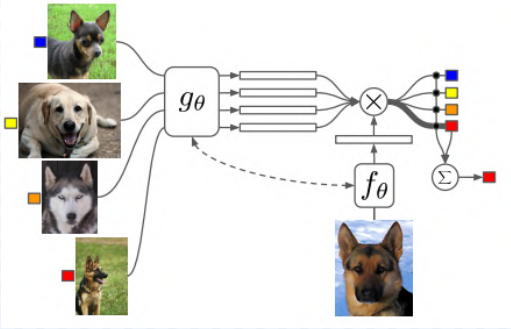
多任务学习

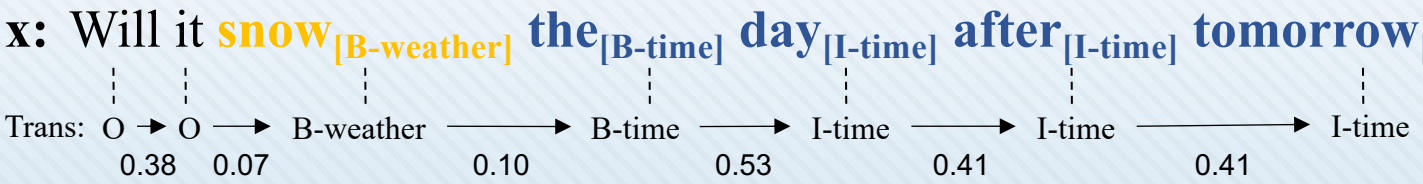
- N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models (Che et al., arXiv:2009.11616 SCIR)
  - 最新版语言技术平台（LTP）
    - https://github.com/HIT-SCIR/ltp/
    - 分词、词性标注、命名实体识别、依存句法分析、语义角色标注、语义依存分析6项任务
  - 基于预训练模型
  - 采用蒸馏多任务学习技术
    - 速度更快
    - 准确率更高
    - 模型更小



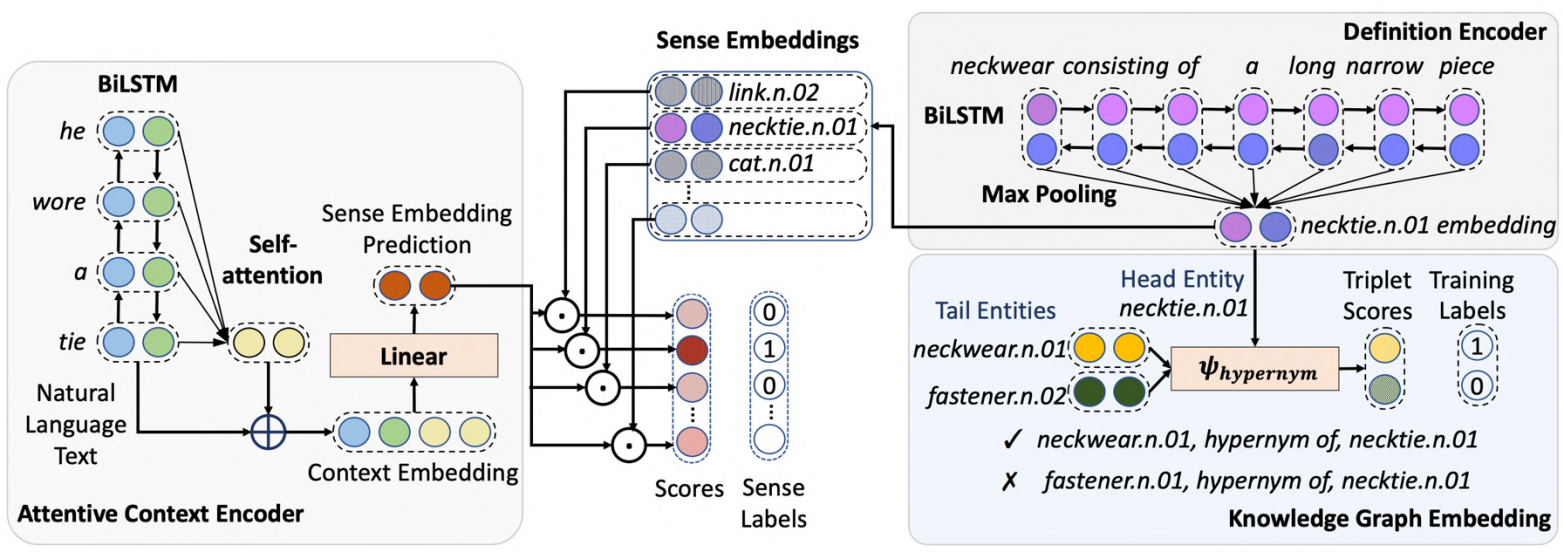
小样本学习

- Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network (Hou et al., ACL 2020 SCIR)
  - 小样本学习目前多应用于分类任务
  - 如何将小样本学习应用于序列标注？
    - 标签之间互相影响，新的领域有新的标签集
  - 利用CRF模型建模
    - 转移概率：提出一种回退机制，建模未见标签的转移概率
    - 发射概率：利用Pair-wise Embedding更好计算词相似度





- Zero-shot Word Sense Disambiguation using Sense Definition Embeddings (Kumar et al., ACL 2019)
  - 上下文词向量与知识图谱词义向量进行比对



GPT-3的应用模式

- GPT-3提出了新的、无需精调的应用模式
  - 给定前文（小样本样例）
  - 继续生成输出结果

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

大纲

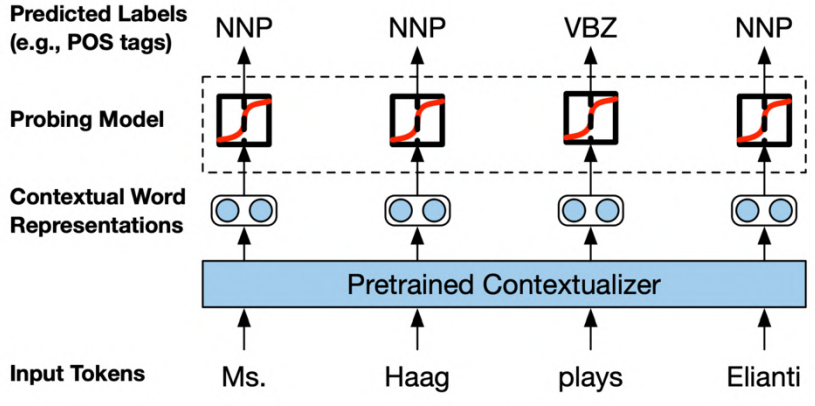
- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

预训练模型分析

- 加入探针（Probe），对模型的性质进行一定的分析
- 增加模型的可解释性，指导设计更好的模型
- 探针的种类
  - 下游任务探针
  - 词向量探针
  - 注意力探针

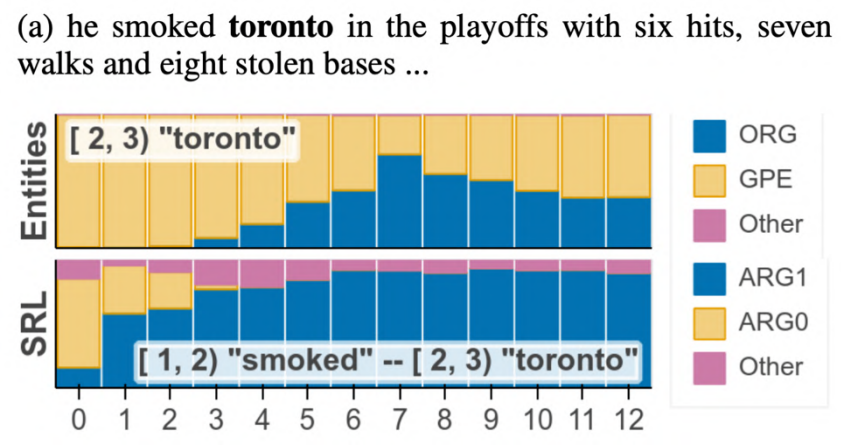
下游任务探针

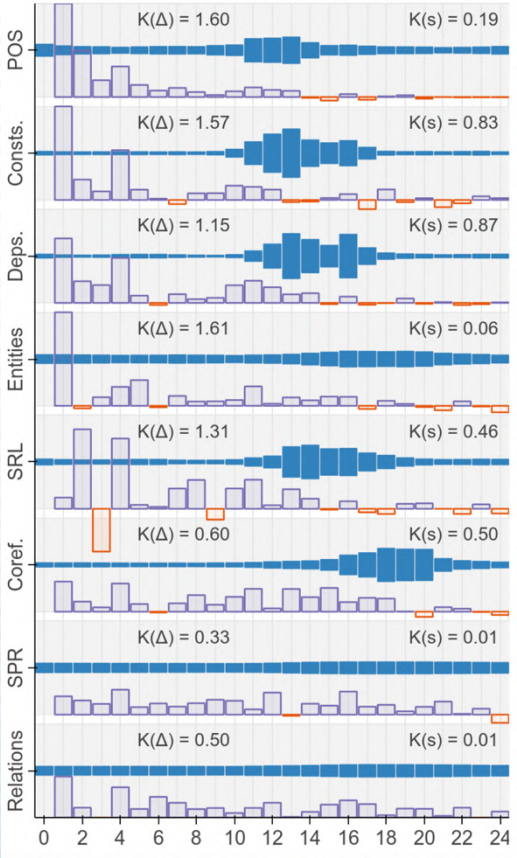
- Linguistic Knowledge and Transferability of Contextual Representations (Liu et al., NAACL 2019)
  - 在16个下游任务中进行实验
    - 固定预训练模型，作为特征提取器
    - 最上层只使用任务相关的线性分类器
  - 结论
    - 预训练模型在大部分任务中表现优异
    - 除了需要细粒度语言知识的任务
      - 如语法检查、NER、并列成分识别等
    - RNN模型（如ELMOs）的上层和任务相关
    - Transformer表现并非如此
    - 在相关有指导任务上预训练，效果比在语言模型上预训练好
    - 随着预训练语言模型数据的增加，其效果越来越好，甚至超过在相关有指导任务上预训练



级联任务探针

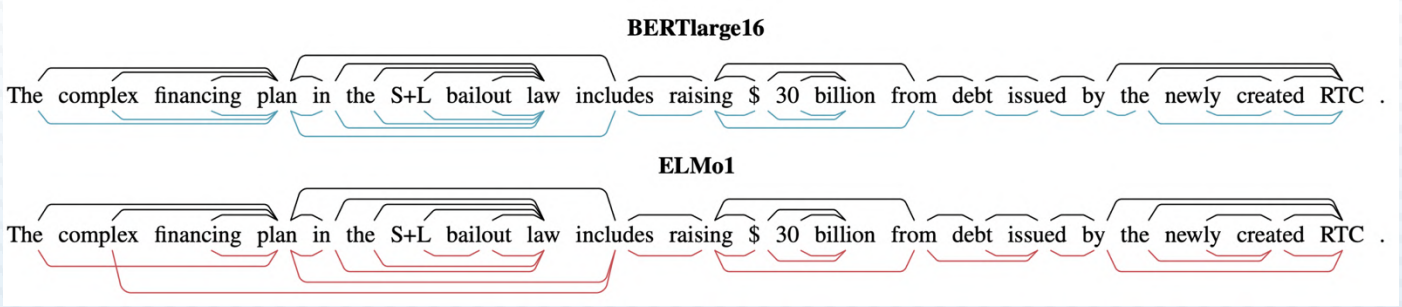
- BERT Rediscovers the Classical NLP Pipeline (Tenney et al., arXiv:1905.05950)
  - 词性标注、短语结构句法分析、依存句法分析、命名实体、语义角色标注、指代消解、关系分类等
  - 和人的直觉类似，这些任务在BERT中是顺序处理的
  - 底层的歧义信息可以通过高层进行调整





句法结构探针

- A Structural Probe for Finding Syntax in Word Representations (Hewitt and Manning, NAACL 2019)
  - 直接计算两个向下文词向量之间的平方距离，最近的画一条弧
  - 预训练上下文词向量蕴含了句子的句法结构信息

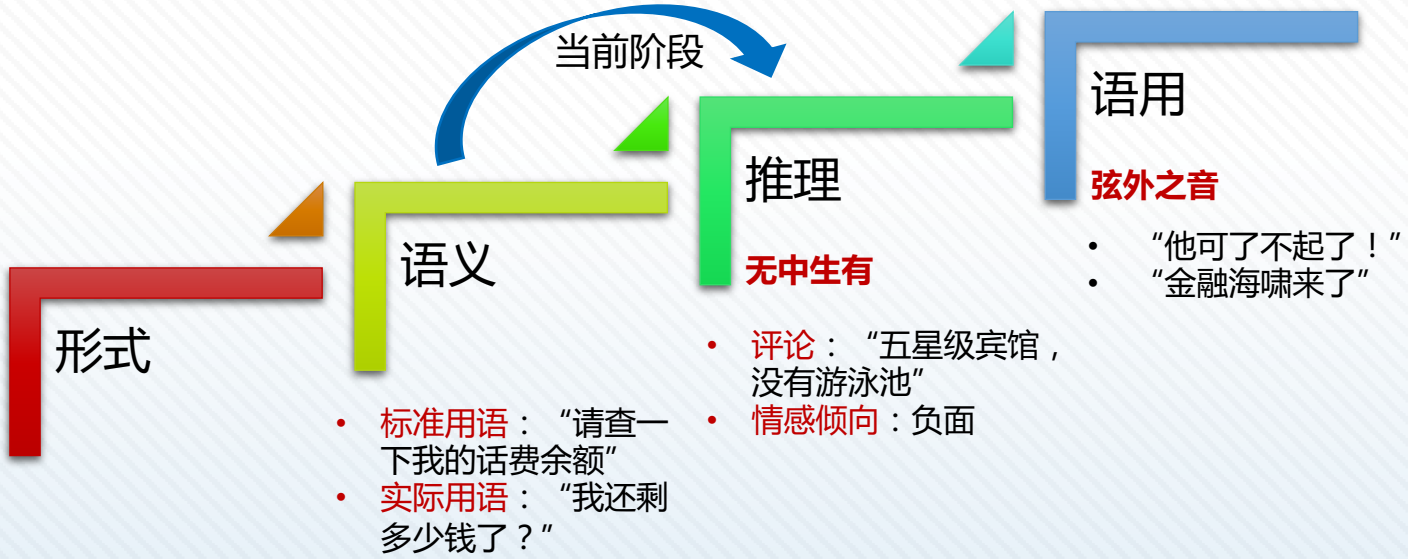


大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 中文预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

开放问题

- 预训练模型是NLP问题的终极解决方案么？



- BERT等预训练模型能很好的解决语义问题
- 但是还无法解决推理问题，因为不是所有的知识都显示的表示在文本中
  - GLUE → SuperGLUE

开放问题

- 预训练模型是NLP问题的终极解决方案么？
- 如何获得更多更好的预训练数据？
  - 伪数据
    - 是带标签的预训练数据
    - 不曾面向所研究的任务进行人工标注
    - 标签是样本的近似答案，而不是精确答案
  - 伪数据的类型
    - 寻“找”自然标注大数据
    - 制“造”标注大数据数据

	任务	方法
修改（换）	词义消歧	等价伪词 (Lu et al., ACL 2006)
删除（挖）	零指代	基于挖词模型 (Liu et al., ACL 2017)
增加（插）	文本顺滑	序列标注 (Wang et al., AAAI 2020)

开放问题

- 预训练模型是NLP问题的终极解决方案么？
- 如何获得更多更好的预训练数据？
- 如何对长文档进行表示？
- 如何解释预训练模型的结果？
- 如何应对攻击？

总结

- 预训练词向量开启了基于深度学习的NLP时代
- 以BERT为代表的预训练模型成为NLP的新范式
- BERT启发了越来越多的预训练模型
- 预训练模型的精调方法及更多应用
- 对预训练模型工作机理的分析
- 预训练模型的研究挑战

感 谢 聆 听

理解语言，认知社会

以中文技术，助民族复兴

长按二维码，关注哈工大SCIR

微信号：HIT_SCIR