预训练模型

自然语言处理的新范式

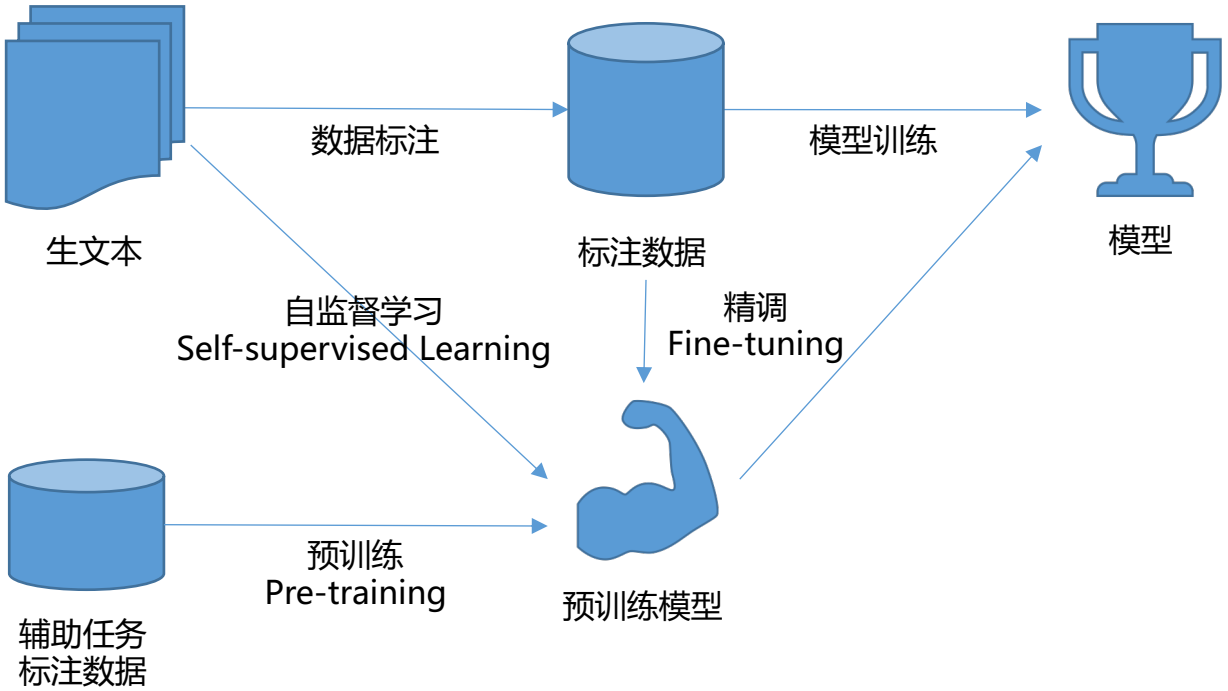
车万翔

社会计算与信息检索研究中心

哈尔滨工业大学

2019-10-18

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
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

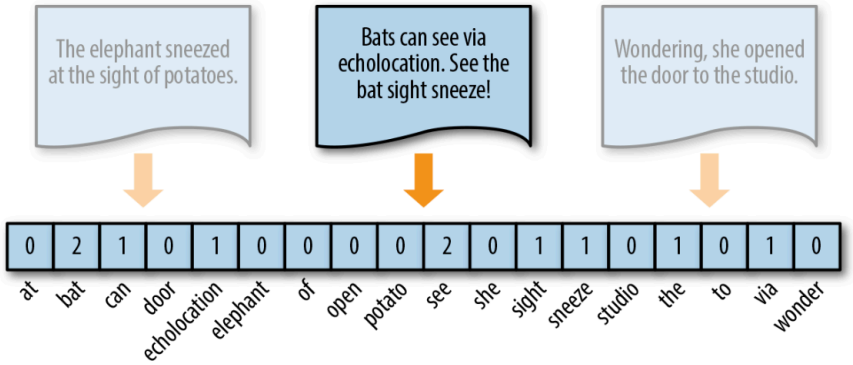
什么是词向量？

- 词的一种机内表示形式，便于计算
- 传统使用one-hot词向量表示词
  - 高维、稀疏、离散
  - 导致严重的数据稀疏问题
  - 所有向量都是正交的

star [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...]

moon [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...]

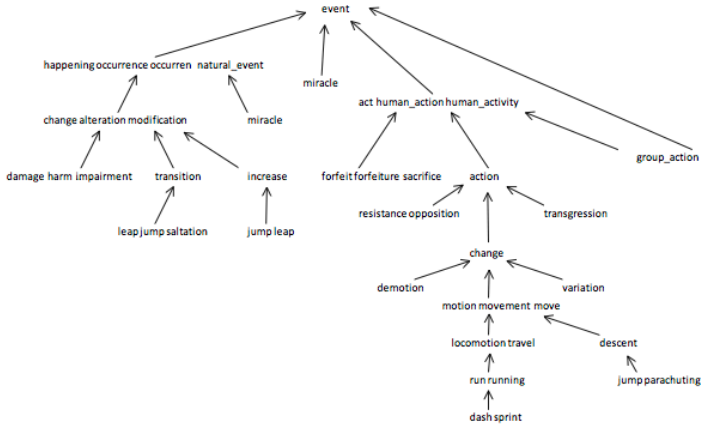
sim(star, moon) = 0 

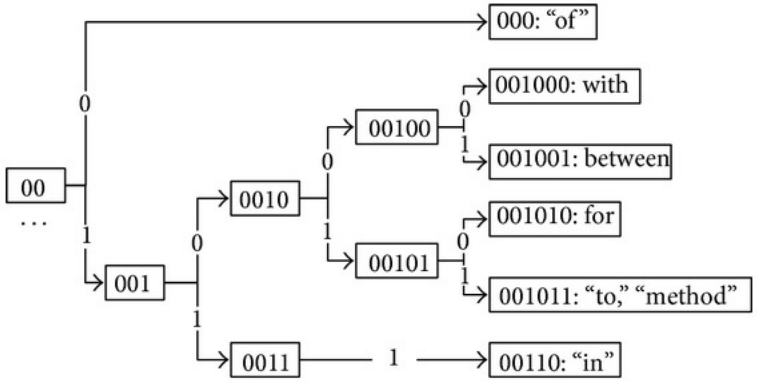


词袋模型 (Bag of Words Model)

传统解决方案

- 增加额外的特征
  - 词性特征：名词、动词、形容词
  - 前后缀特征：re-、-tion、-er
- 语义词典
  - WordNet、HowNet等
  - 如词的上位信息表示语义类别
  - 需要解决一词多义问题
  - 收录的词不全且更新慢
- 词聚类特征
  - 如Brown Clustering (Brown et al., CL 1992)





词的分布语义假设

- 分布语义假设（distributional semantic hypothesis）
  - 词的含义可由其上下文词的分布进行表示
    - *You shall know a word by the company it keeps* -- Firth J.R. 1957

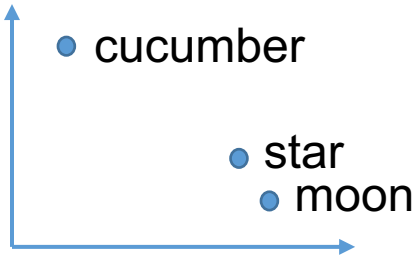
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind

词的分布（Distributional）表示

- 分布词向量

	shinning	bright	trees	dark	look
moon	38	45	2	27	12

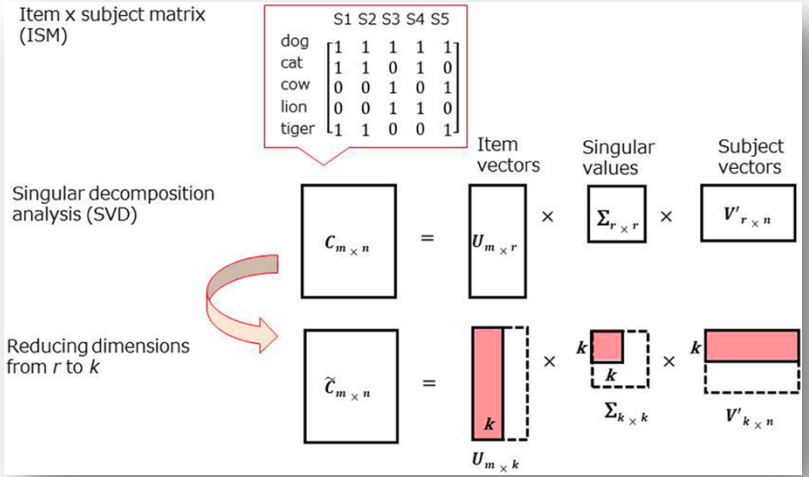
- 语义相似度通过计算向量相似度获得



- 仍然存在高维、稀疏、离散的问题

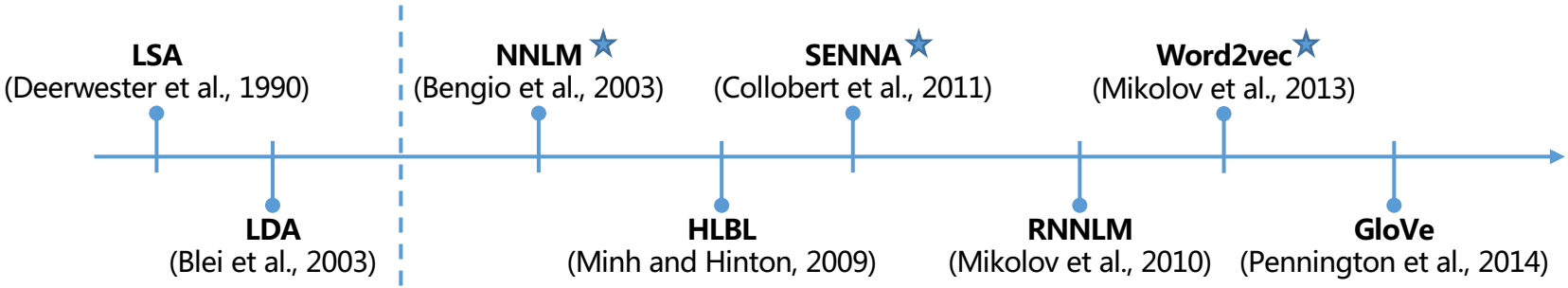
分布表示的优化及优缺点

- 高维、稀疏、离散 ➔ 低维、稠密、连续
  - 加权
    - TF-IDF
    - PMI (Pointwise Mutual Information)
  - 降维
    - Singular Value Decomposition (SVD)
    - Latent Dirichlet Allocation (LDA)
- 优缺点
  - 优点
    - 容易实现，可解释性强
  - 缺点
    - 训练速度慢，增加新语料库困难
    - 不易扩展到短语、句子表示



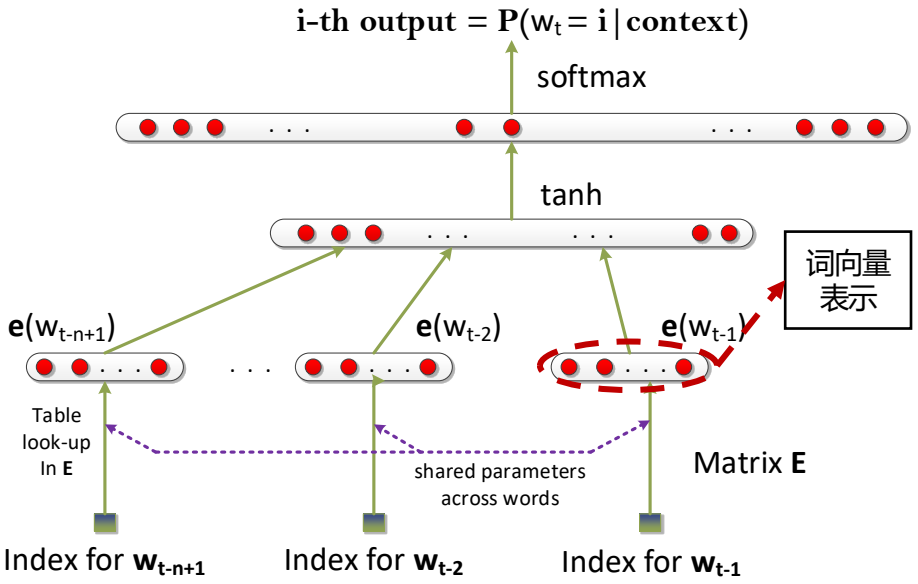
分布式（Distributed）词表示

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



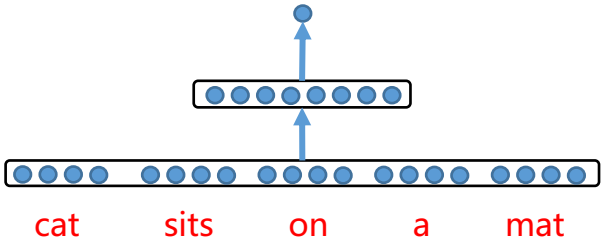
SENNA

- 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
  - 可利用大规模数据
  - 弥补了模型能力的不足

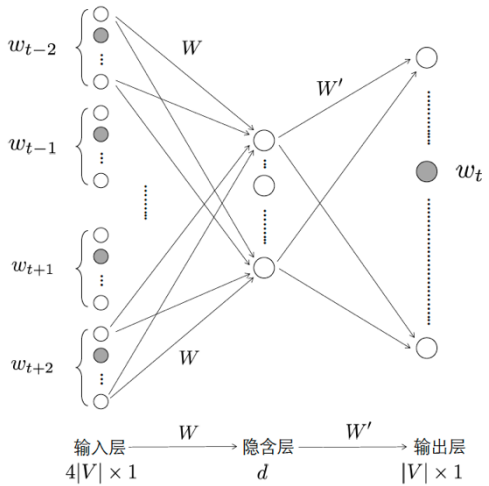


Figure 3: CBOW模型。

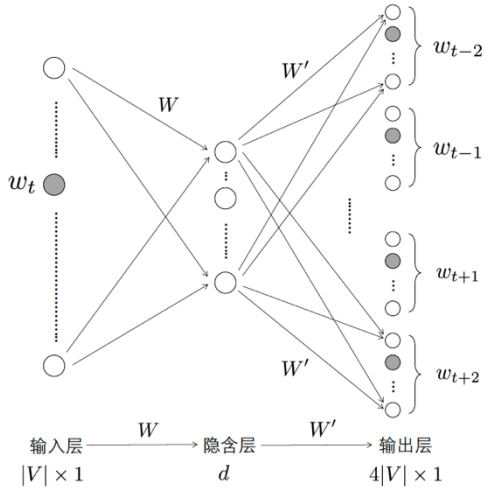
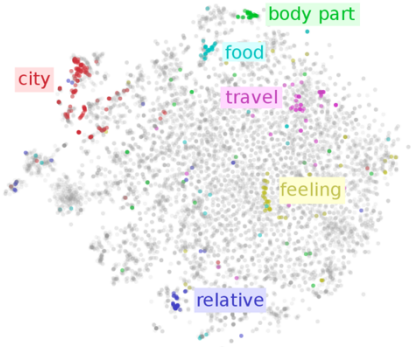
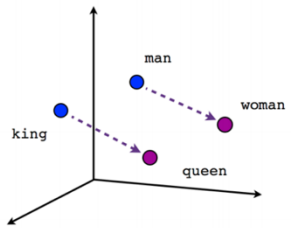


Figure 4: Skip-gram模型。

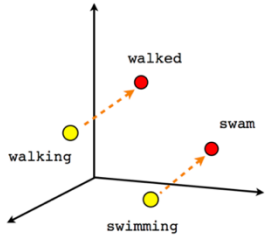
词向量的应用



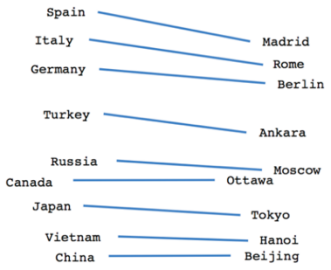
词义相似度计算



Male-Female

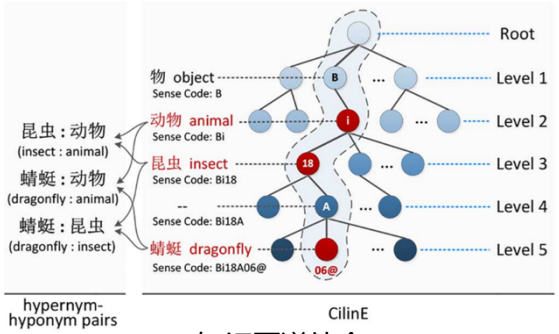


Verb tense

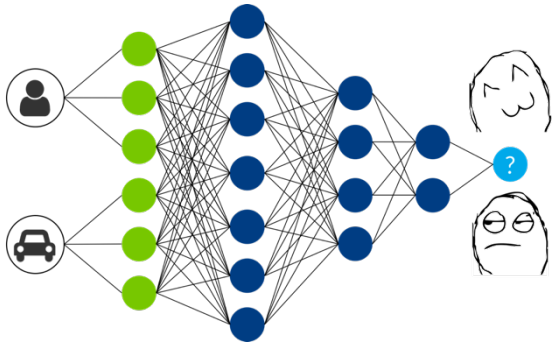


Country-Capital

词类比关系计算



知识图谱补全



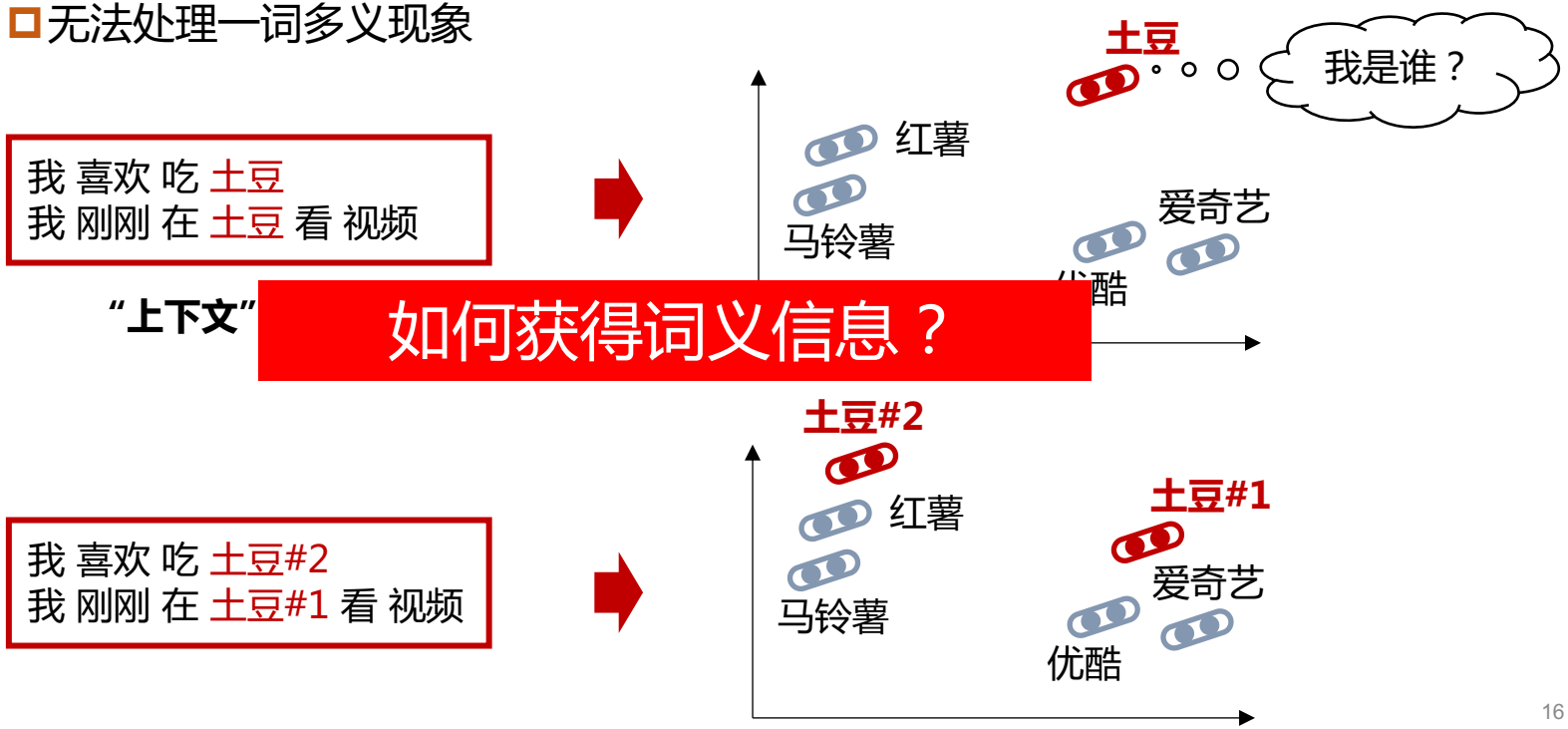
推荐系统

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

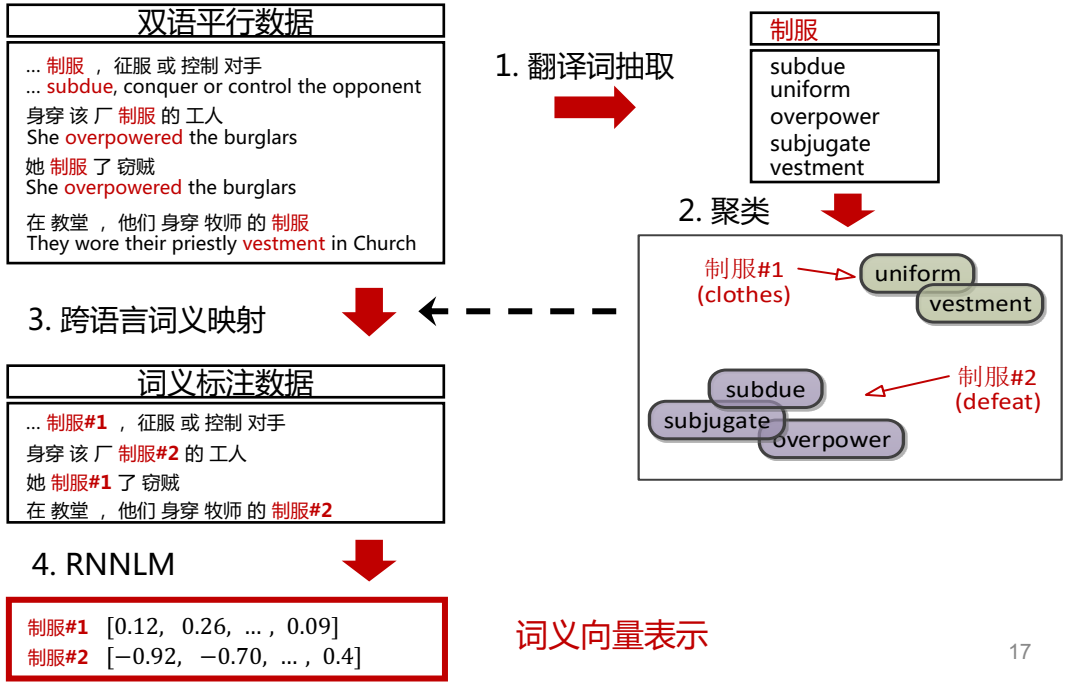
一词多义现象

- 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



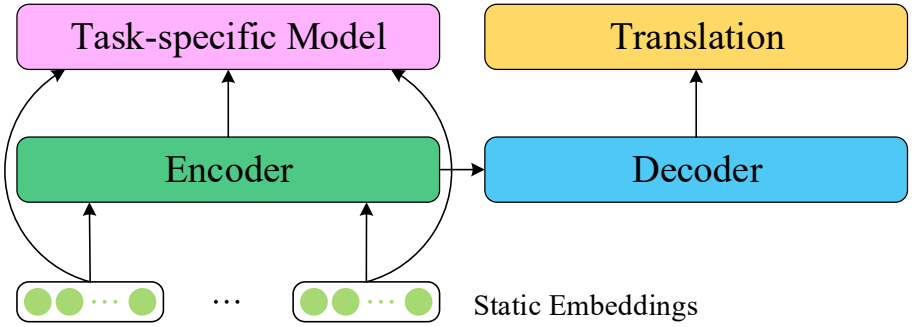
基于双语的词义向量表示

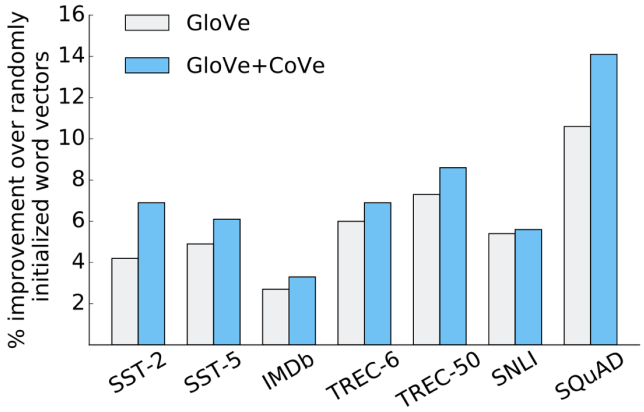
- Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources (Guo et al., Coling 2014)



CoVe

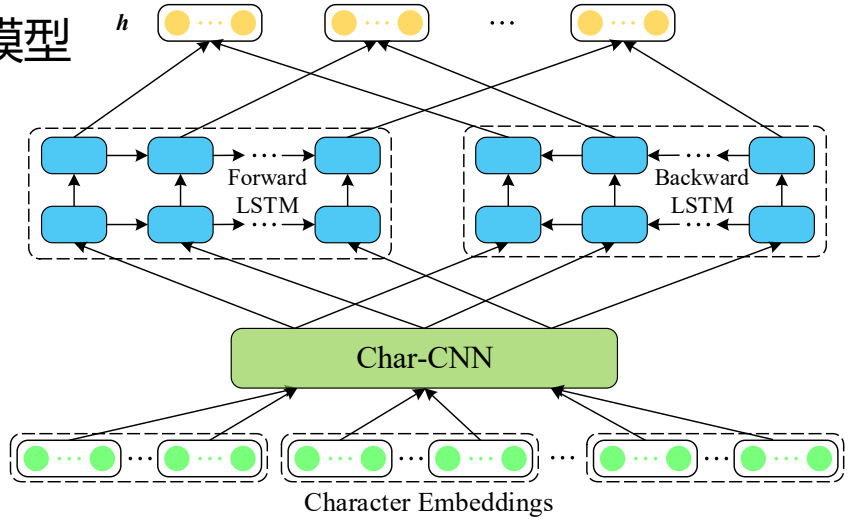
- Learned in Translation: Contextualized Word Vectors (McCann et al., arXiv:1708.00107)
  - CoVe: Context Vectors
- 预训练NMT模型
- 将Encoder作为目标任务的额外特征



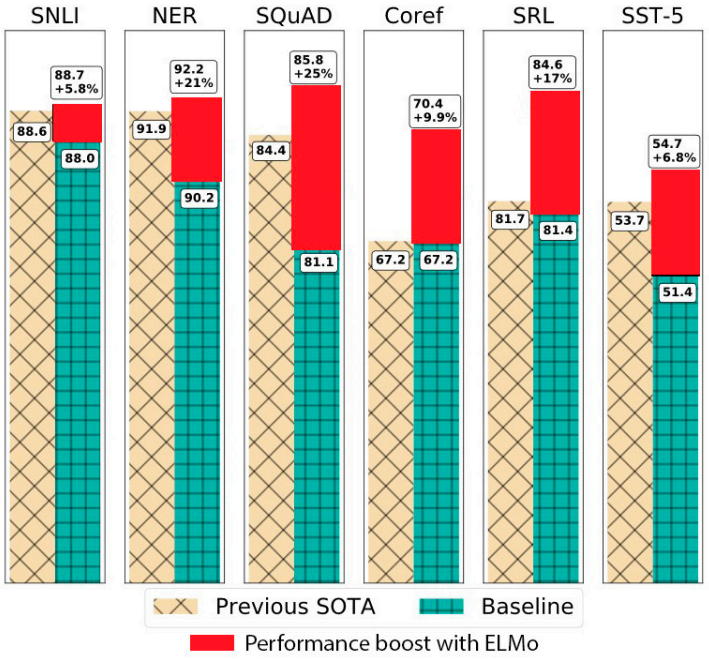


ELMo

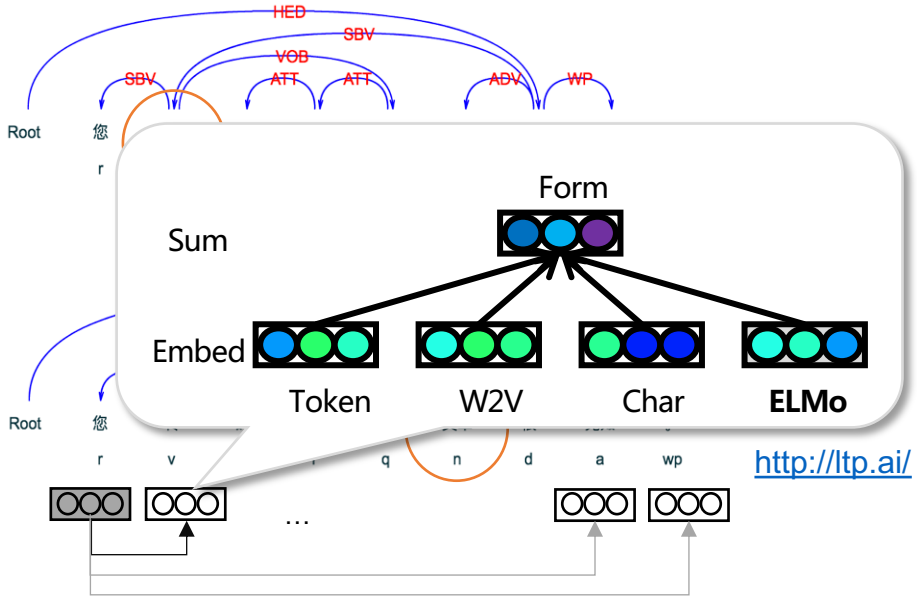
- Deep Contextualized Word Representations (Peters et al., NAACL 2018)
  - ELMo: Embeddings from Language Models
- 使用字符的CNN表示词
- 分别训练从左至右和从右至左的语言模型
- 使用语言模型的输出作为词向量特征
- 语言模型训练数据接近“无限”



基于ELMo的应用



- 依存句法分析 (Che et al., CoNLL 2018)

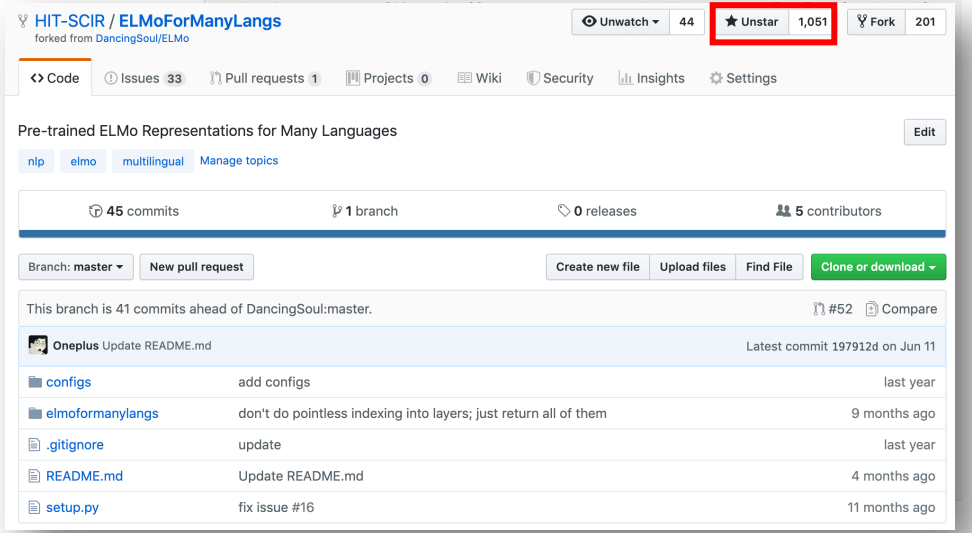


CoNLL 2018评测

- 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

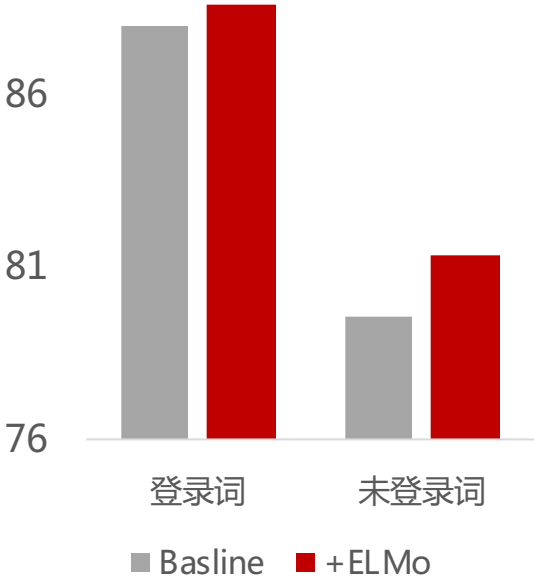
LAS Ranking

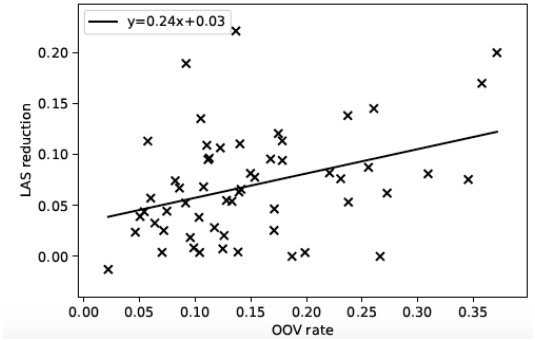
1.	HIT-SCIR (Harbin)	75.84 ± 0.14 [OK] (p<0.001)
2.	TurkuNLP (Turku)	73.28 ± 0.14 [OK] (p=0.039)
3-5.	UDPipe Future (Praha)	73.11 ± 0.13 [OK] (p=0.221)
3-5.	LATTICE (Paris)	73.02 ± 0.14 [OK] (p=0.461)
3-5.	ICS PAS (Warszawa)	73.02 ± 0.14 [OK] (p<0.001)
6.	CEA LIST (Paris)	72.56 ± 0.14 [OK] (p=0.036)
7-8.	Uppsala (Uppsala)	72.37 ± 0.15 [OK] (p=0.191)
7-8.	Stanford (Stanford)	72.29 ± 0.14 [OK] (p<0.001)



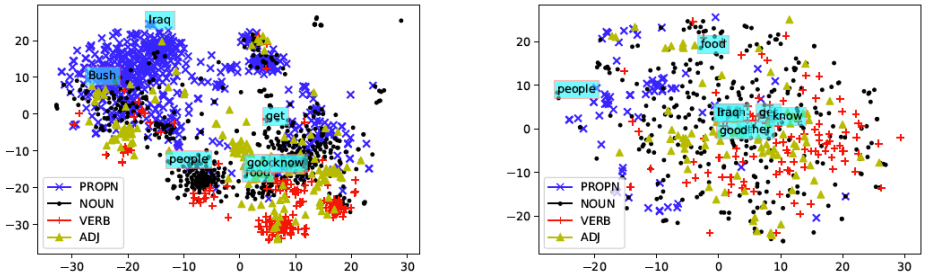
ELMo为什么有效？

- 有效提升未登录词的准确率 (Liu et al., TALLIP 2019)





ELMo带来的性能提升与未登录词比例正相关



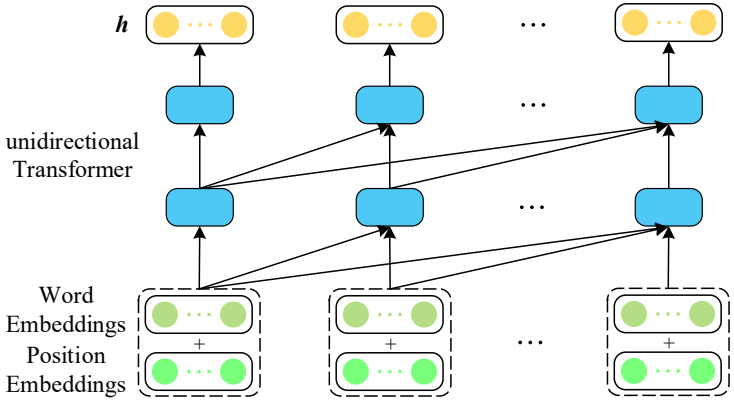
未登录词的可视化
（左：上下文相关词向量，右：Word2vec）

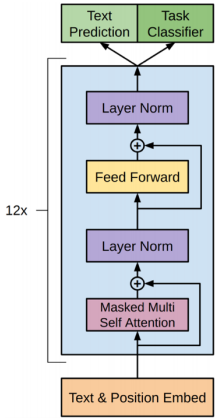
大纲

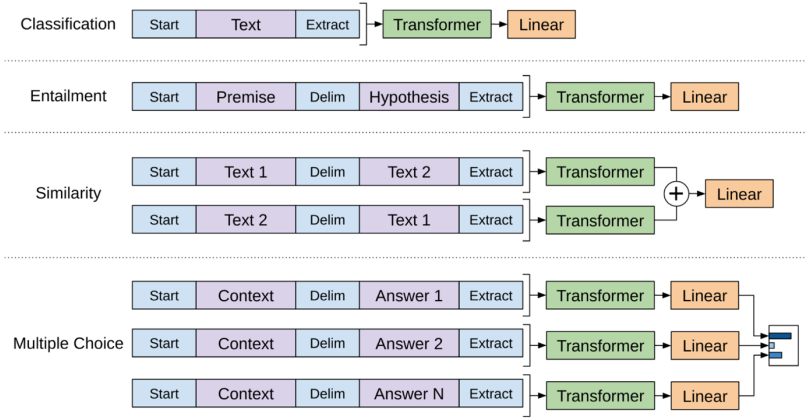
- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

GPT

- 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

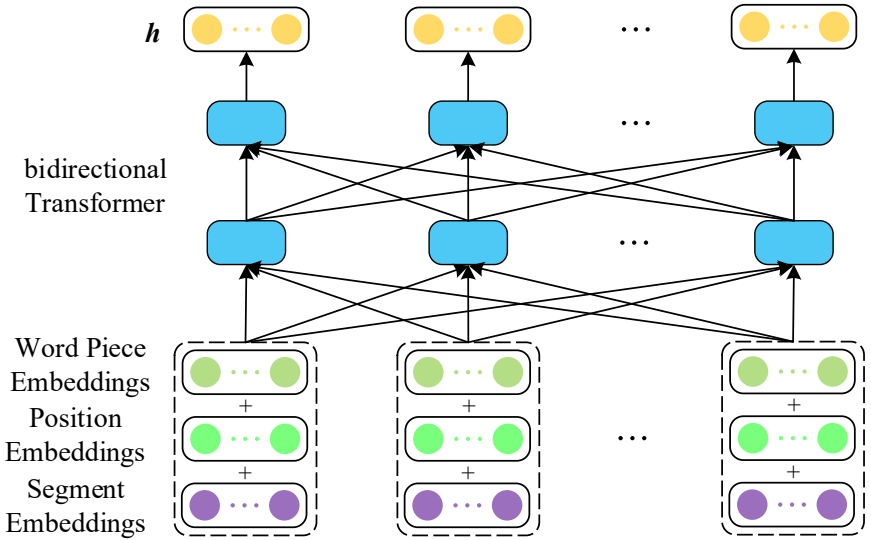






BERT

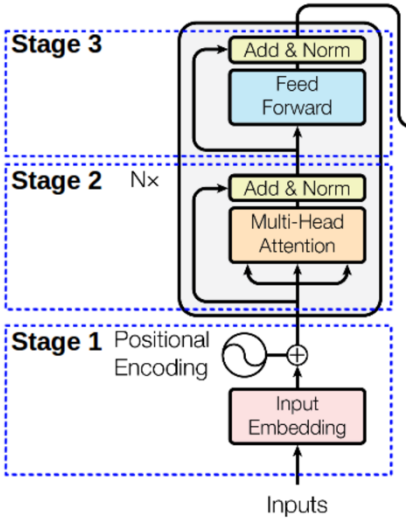
- Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019)
  - BERT: **Bidirectional** Encoder Representations from Transformers



BERT模型详解

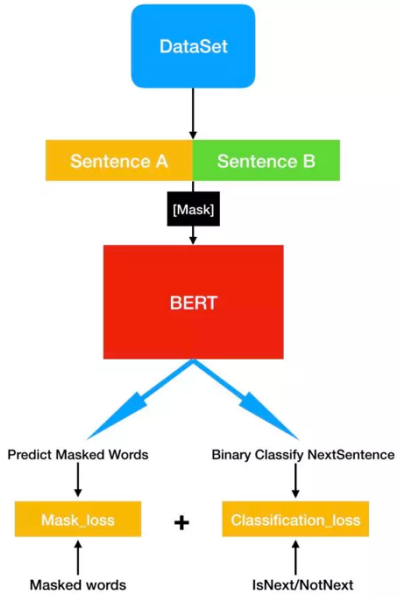
编码器

- 输入：Word Piece
- 编码器：Transformer



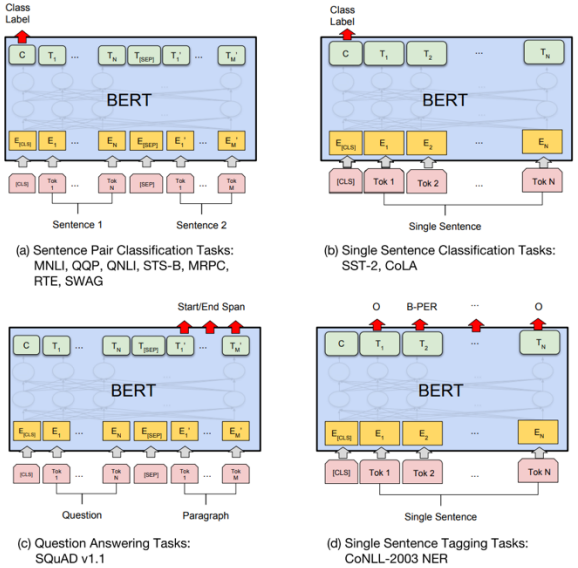
预训练任务

- 完形填空 + 下句预测 (NSP)



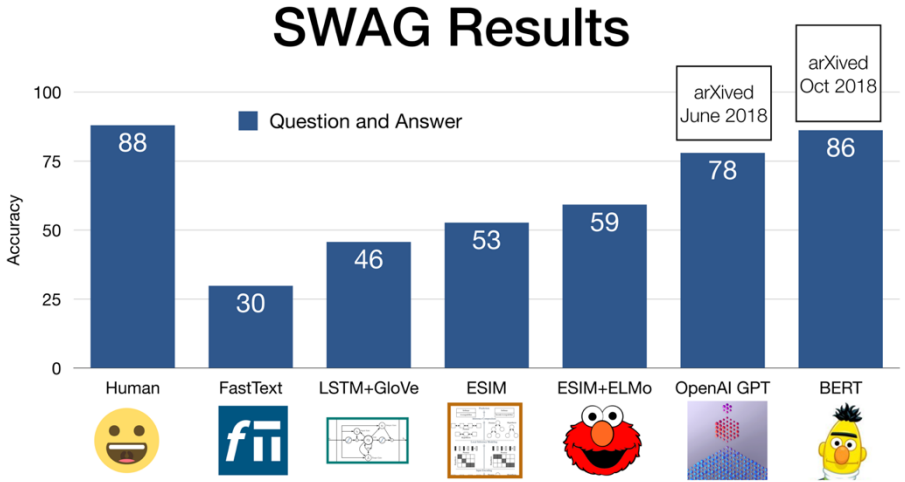
应用方式

- 在目标任务上Fine-tune
- 四种任务类型



BERT的应用效果

- 论文中刷新了 11 项 NLP 任务的当前最优性能记录
- 后续工作表明其显著提高了众多其它任务性能

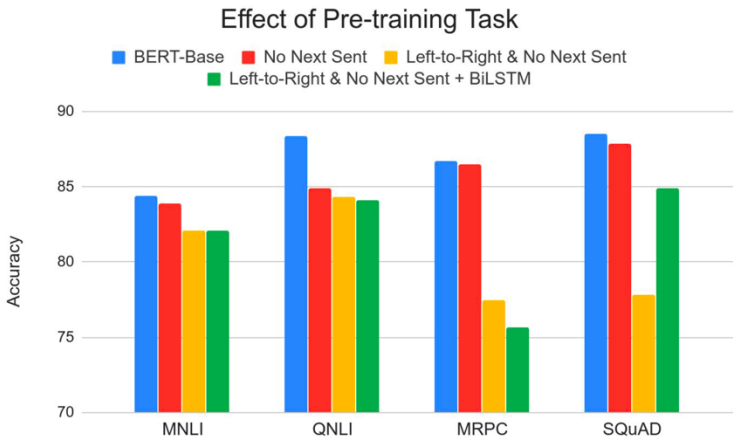


Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
5 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204
5 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
6 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967

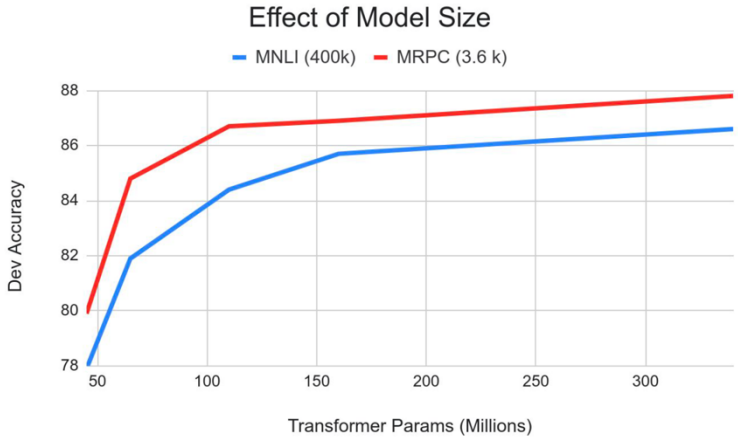
SQuAD2.0
The Stanford Question Answering Dataset

BERT中各种策略的影响

预训练任务



模型大小



BERT改进模型

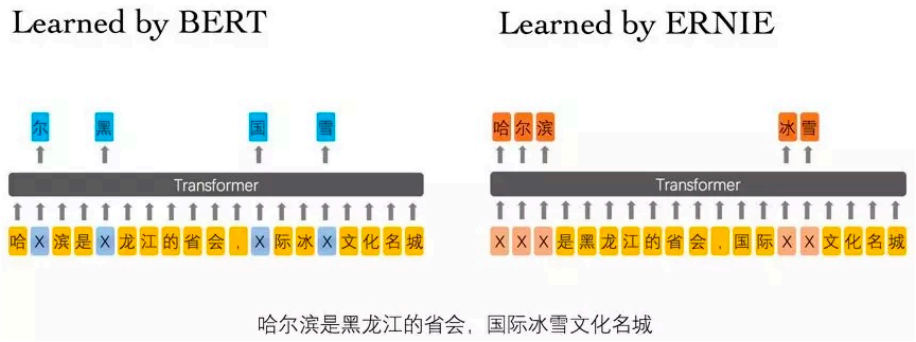
- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 跨语言与跨模态

BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 跨语言与跨模态

ERNIE（百度）

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

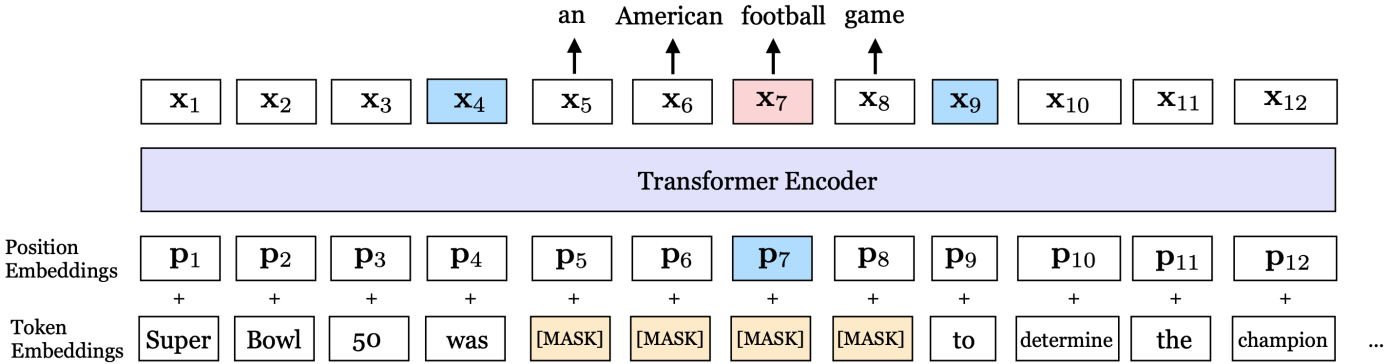


任务	ERNIE 1.0 模型	ERNIE 2.0 英文模型	ERNIE 2.0 中文模型
Word-aware	Knowledge Masking	Knowledge Masking Capitalization Prediction Token-Document Relation Prediction	Knowledge Masking
Structure-aware		Sentence Reordering	Sentence Reordering Sentence Distance
Semantic-aware	Next Sentence Prediction	Discourse Relation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

SpanBERT

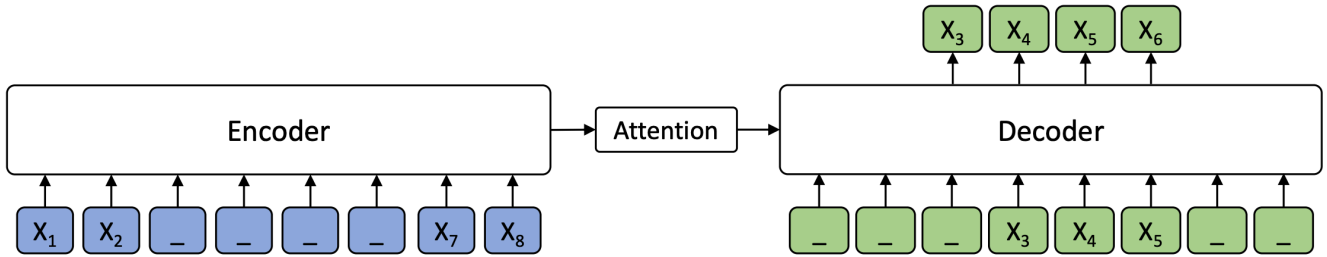
- SpanBERT: Improving Pre-training by Representing and Predicting Spans (Joshi et al., arXiv:1907.10529)
  - 挖掉一段文字，通过学习段的边界表示预测段中每个词
  - 去除NSP预训练目标（由于主题不同，容易判断）
  - 在段抽取任务，如抽取式问答中表现良好

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\mathbf{x}_7) + \mathcal{L}_{\text{SBO}}(\mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_7)$$



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

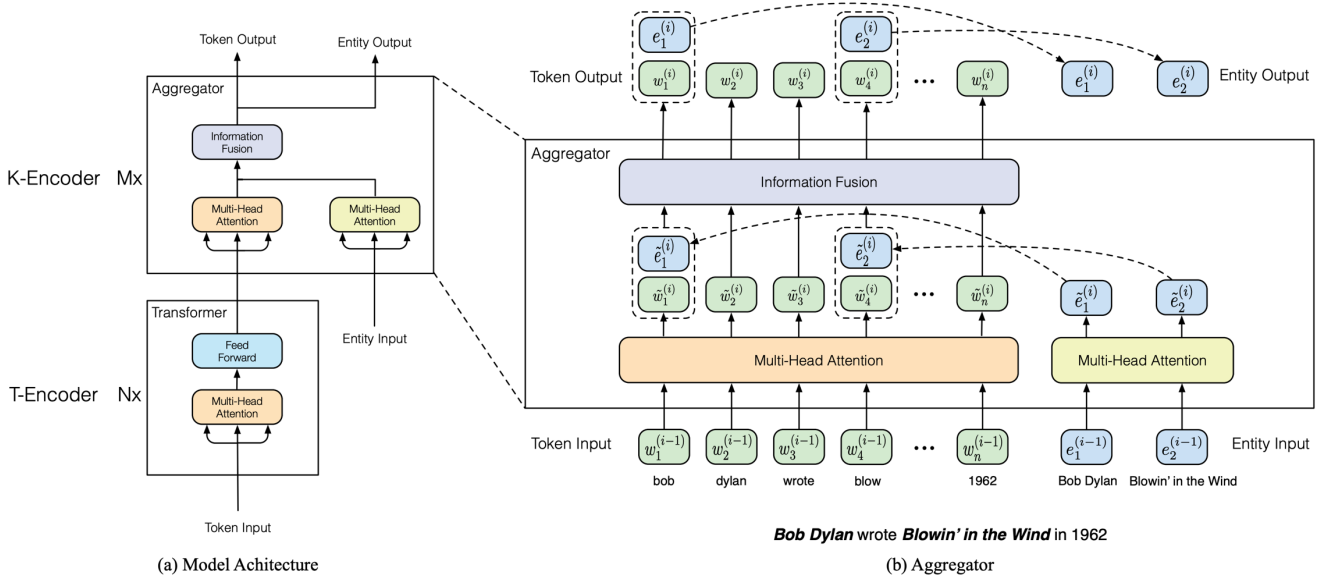


BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 跨语言与跨模态

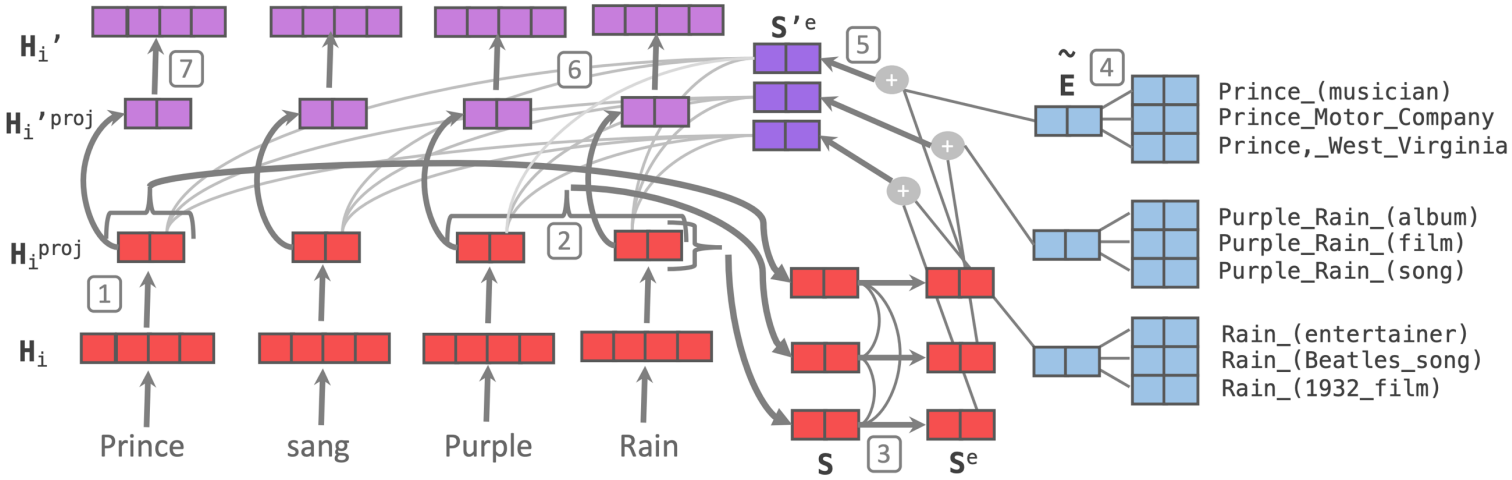
ERNIE（清华）

- 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



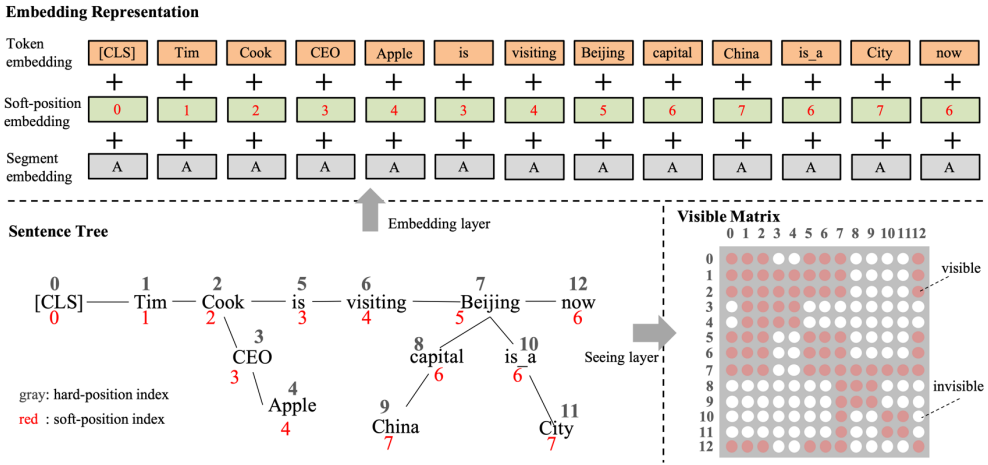
KnowBERT

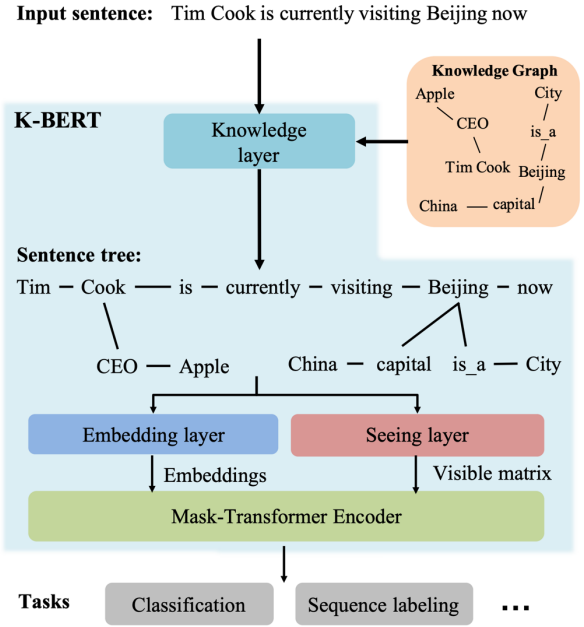
- Knowledge Enhanced Contextual Word Representations (Peters et al., EMNLP 2019)
  - 在融入知识图谱的表示时，使用注意力机制建模交互信息



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





BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 跨语言与跨模态

RoBERTa

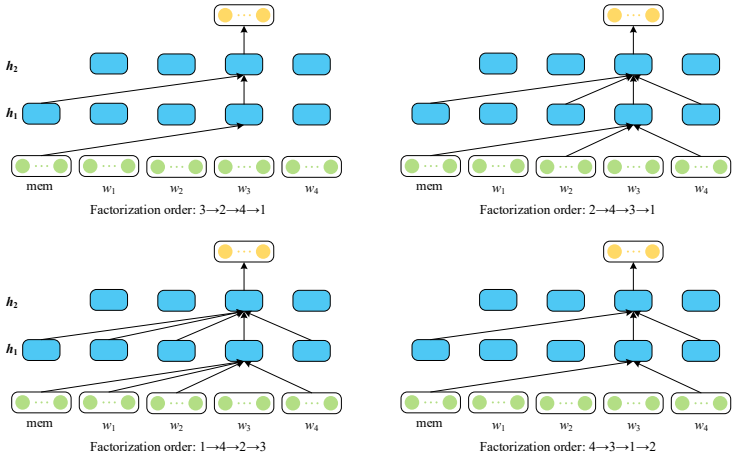
- RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., arXiv:1907.11692)
  - 基于BERT进行细致调参
    - 更多的数据，更大的batch，更长的训练时间
    - 去除NSP任务
    - 训练数据序列更长
    - 训练过程中，动态改变Mask的内容
  - 在1,024块V100 GPU上训练了一天！！

BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 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



BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 跨语言与跨模态

DistilBERT

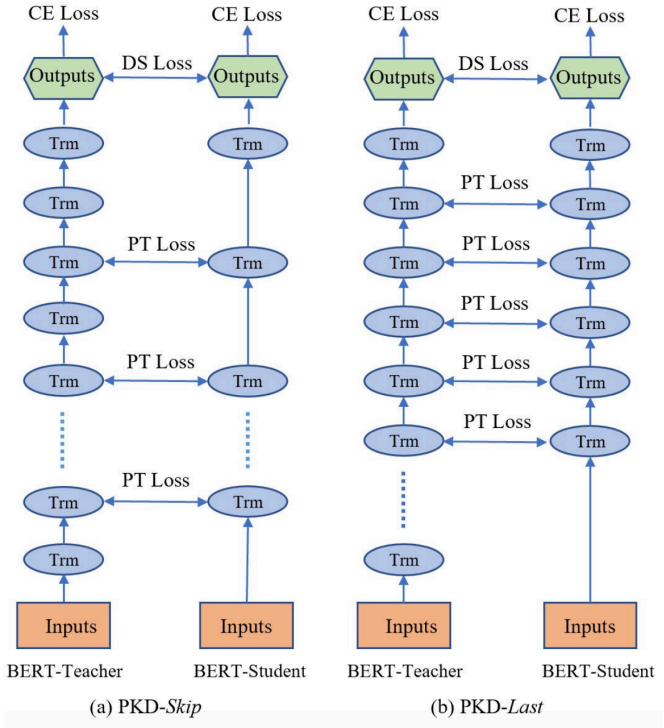
- Distilling BERT (Sanh et al., NeurIPS Workshop 2019)
  - 蒸馏：使用小模型，模仿大模型的预测结果

	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

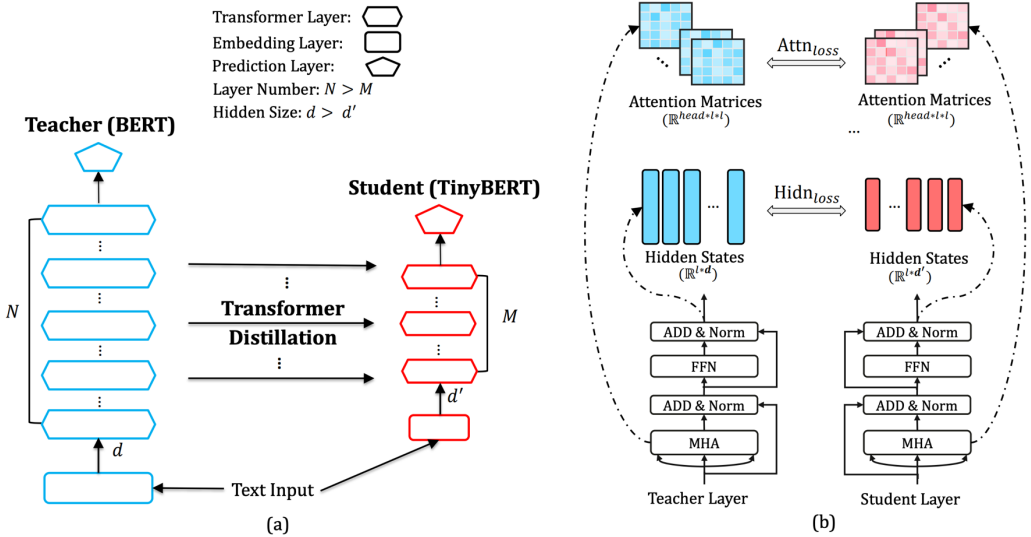
PKD for BERT

- Patient Knowledge Distillation for BERT Model Compression (Sun et al., arXiv:1908.09355)
  - 基于知识蒸馏
    - 按层蒸馏：不只模拟输出层
    - 跳层蒸馏：进一步减小参数量
  - 准确率有一定的降低



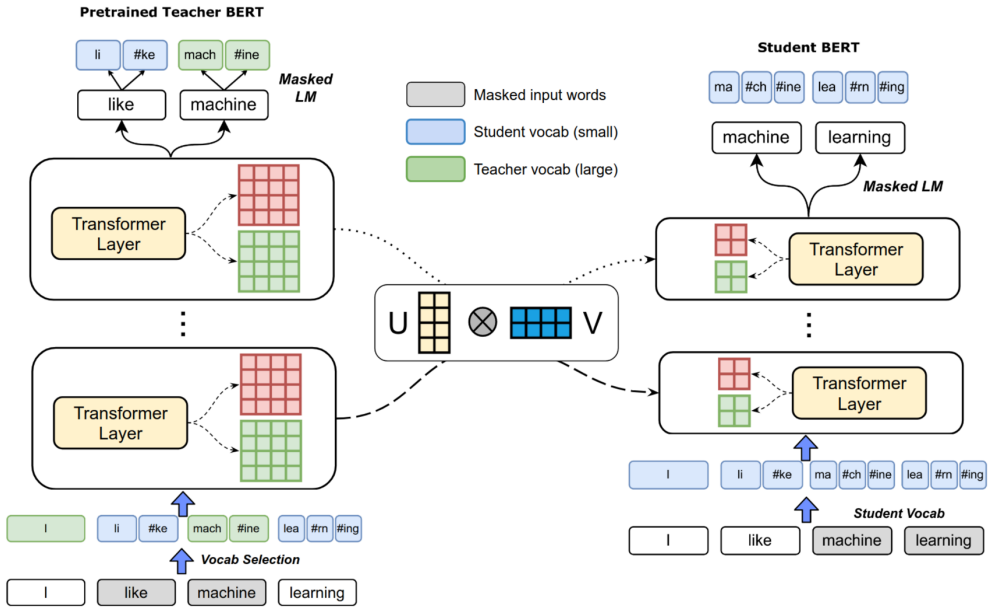
TinyBERT

- TinyBERT: Distilling BERT for Natural Language Understanding (Jiao et al., arXiv:1909.10351)
  - 基于知识蒸馏
  - 学习目标：Teacher模型的
    - 隐层激活
    - 注意力矩阵
  - 最高压缩7.5倍
  - 推理速度快9.4倍
  - 准确率有一定的降低



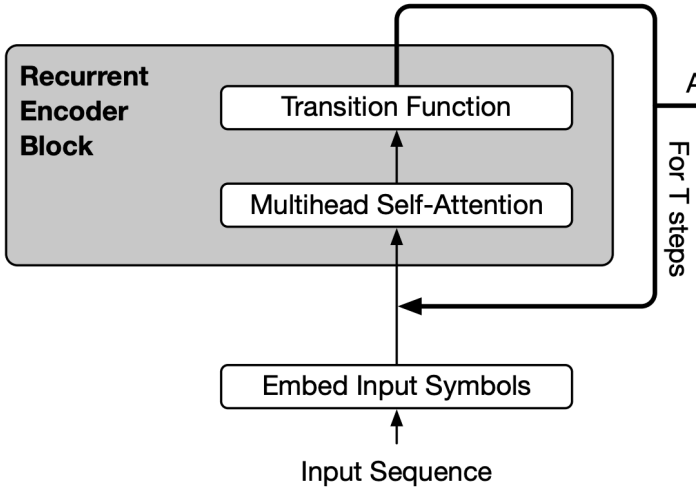
词表优化与逐层映射

- Extreme Language Model Compression with Optimal Subwords and Shared Projections (Zhao et al., arXiv:1909.11687)
  - 基于知识蒸馏
  - 减小词表（30K→5K）
  - 逐层映射（共享映射函数）
  - 最高压缩60倍
  - 准确率有一定的降低



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（128）
  - 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
    - 模型泛化能力有所提高
    - 在多个评测排行榜中位列第一

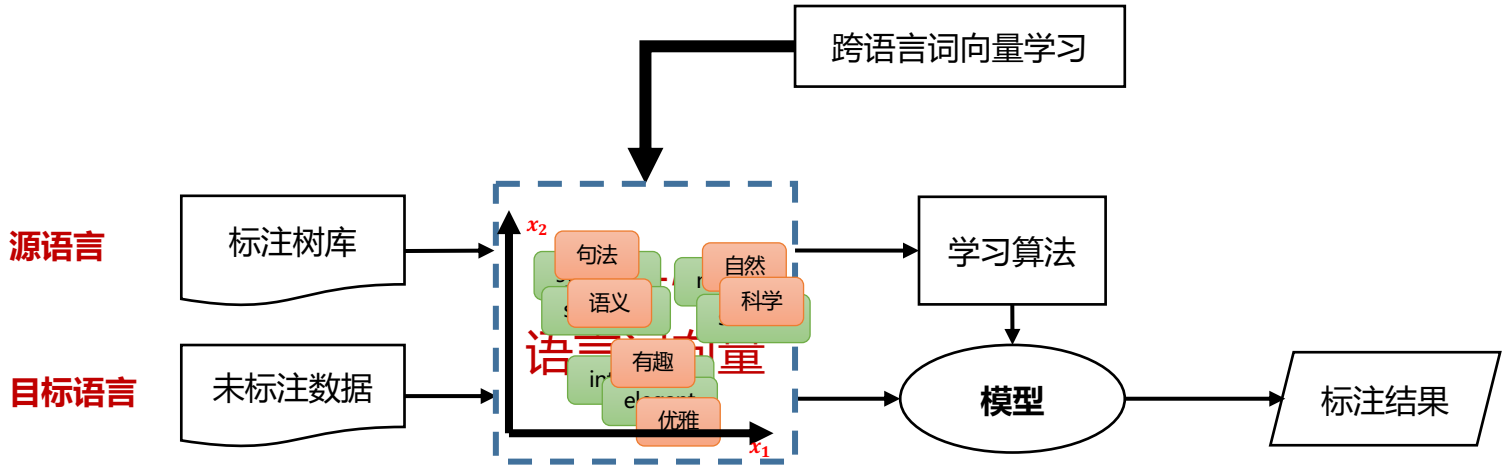


BERT改进模型

- 使用其它预训练目标
- 融入知识图谱
- 更精细的调参
- 解决输入不一致问题
- 模型压缩与加速
- 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)
    - 基于“静态”词向量

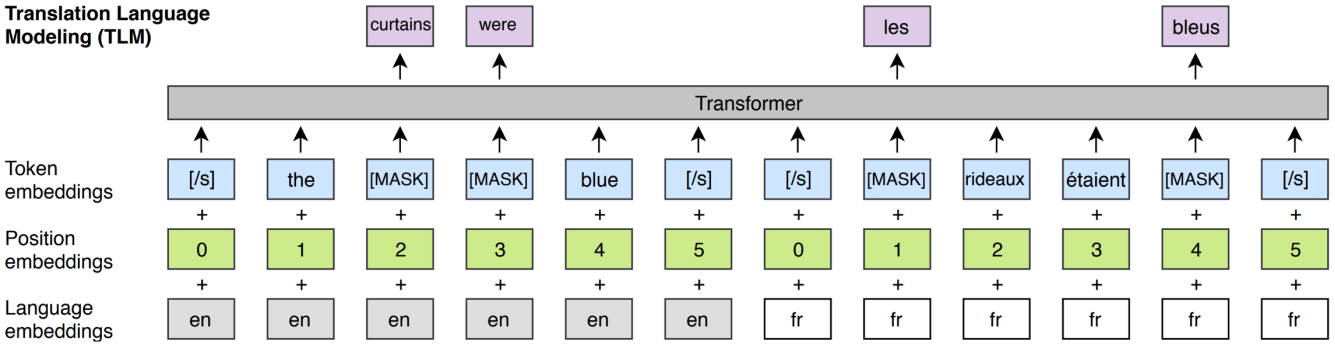


多语言BERT

- 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

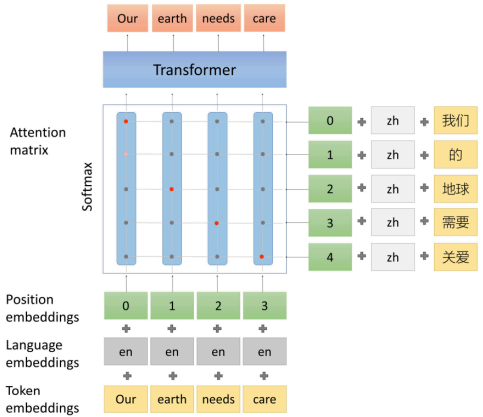
跨语言预训练语言模型

- 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

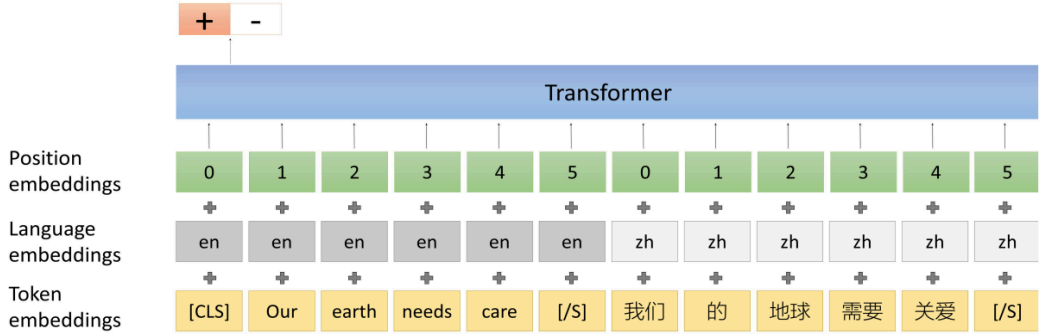


Unicoder

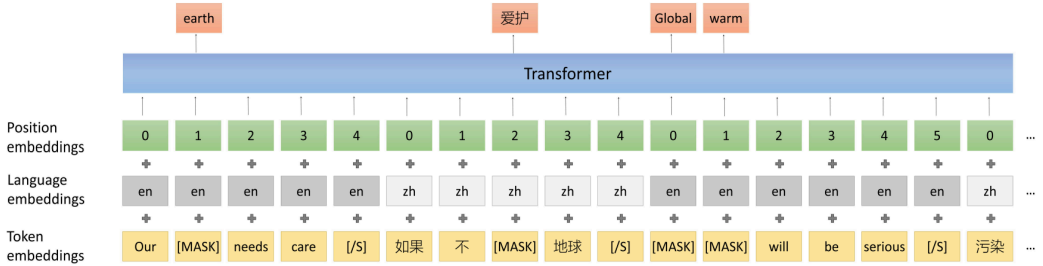
- Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks (Huang et al., EMNLP 2019)
  - 三种跨语言预训练任务
    - (a) 跨语言的词语恢复
    - (b) 跨语言的同义句子分类
    - (c) 跨语言的遮盖语言模型



(a) Cross-lingual Word Recovery



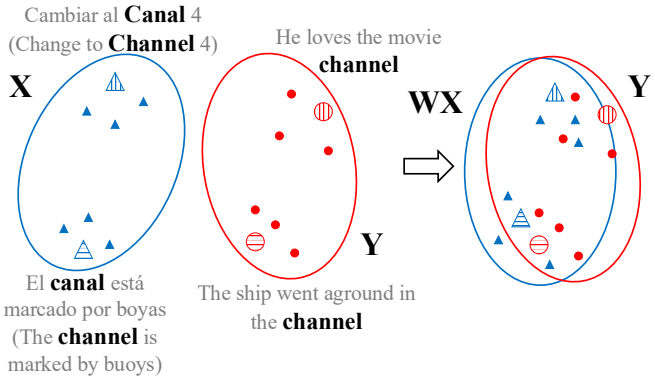
(b) Cross-lingual Paraphrase Classification



(c) Cross-lingual Masked Language Model

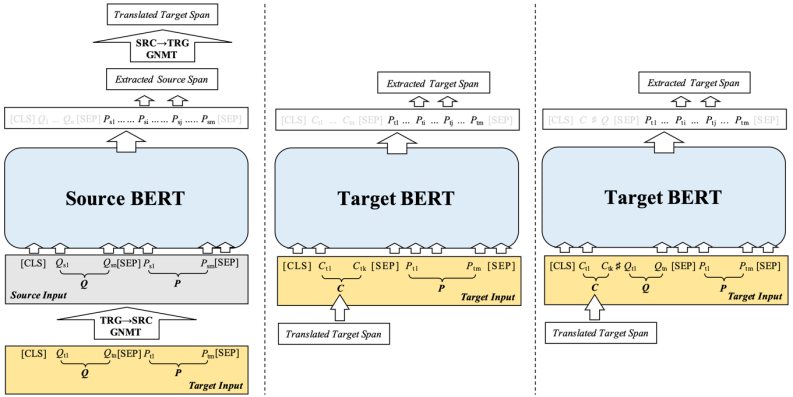
跨语言映射BERT

- Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing (Wang et al., EMNLP 2019)
  - 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

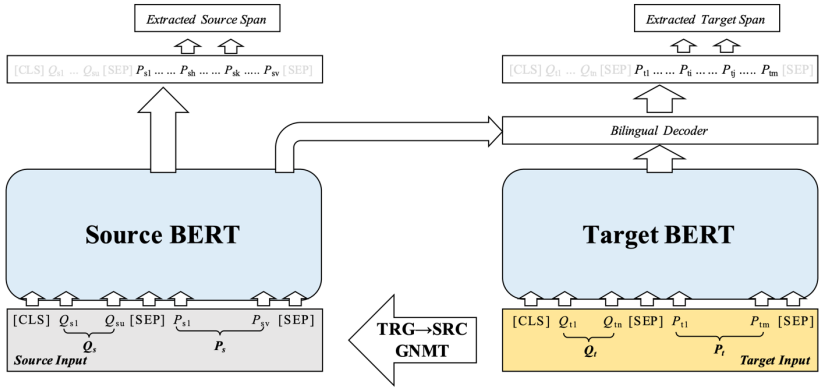


跨语言阅读理解

- Cross-Lingual MRC (Cui et al., EMNLP 2019)
  - 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



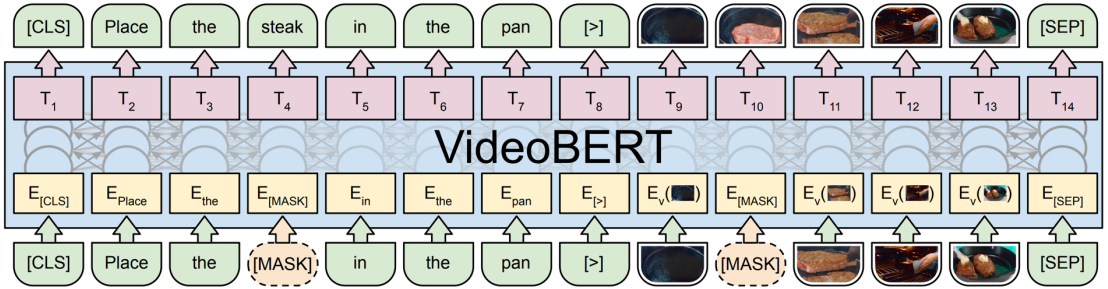
改进回翻技术

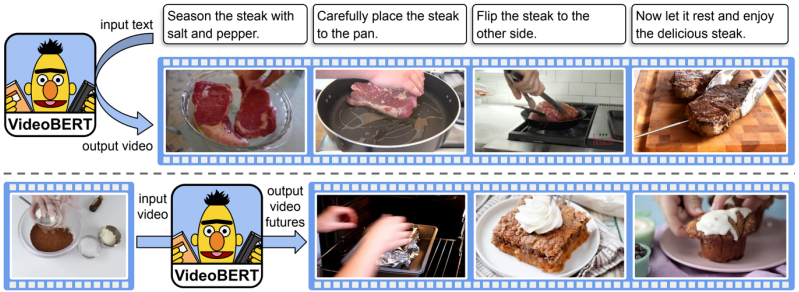


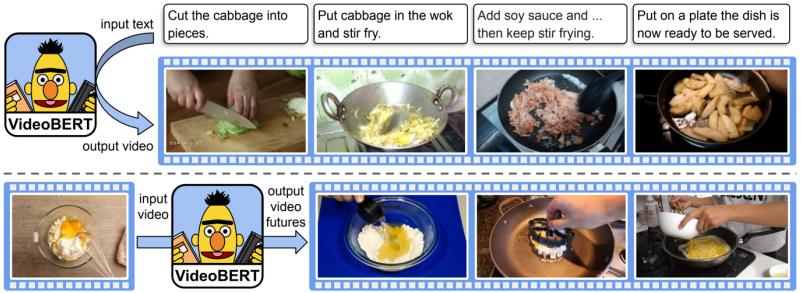
Dual BERT

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







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).

BERT模型改进方法总结

策略	模型	核心技术
使用其它预训练目标	ERNIE 1.0（百度）	Mask中文词或实体
	ERNIE 2.0（百度）	使用词、语义、结构等更多的预训练目标
	SpanBERT	Mask一段文本，并利用段边界的表示预测段中的每个词
	MASS	Mask一段文本，并利用其余文本生成该段文本
融入知识图谱	ERNIE（清华）	将知识图谱中实体的表示融入预训练模型的文本表示
	KnowBERT	在融入知识图谱的实体表示时，使用注意力机制建模交互信息
	K-BERT	在推理阶段融入知识图谱中相关实体和关系的文本表示
更精细的调参	RoBERTa	去掉NSP目标，并调整各种预训练的参数
解决输入不一致问题	XLNet	使用排列语言模型解决输入不一致问题；使用Transformer-XL建模更长的序列
模型压缩与加速	DistilBERT++	使用知识蒸馏技术，以小模型拟合大模型的概率输出结果
	ALBERT	将NSP目标，改为SOP提高了性能；使用参数共享策略和减小词向量维度来压缩模型
跨语言与跨模态	M-BERT	多语言文本同时与训练，共享的词表以及Code-switching起到跨语言效果
	XLM	将双语句对作为BERT的输入，同时Mask双语词
	BERT-Trans	通过线性变换将一种语言的BERT映射为另一种语言
	VideoBERT	将文本和视频对作为BERT的输入，同时Mask词以及图像块

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

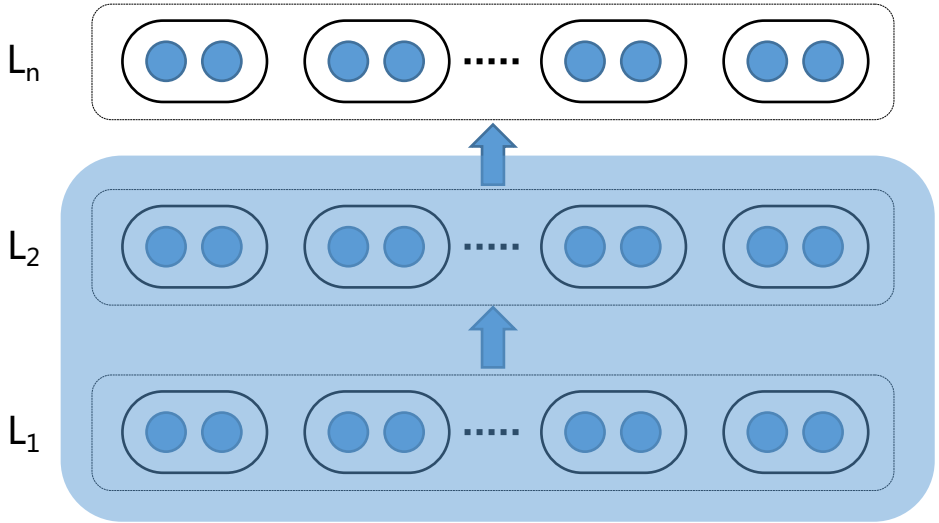
是否需要精调（Fine-tune）？

- 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

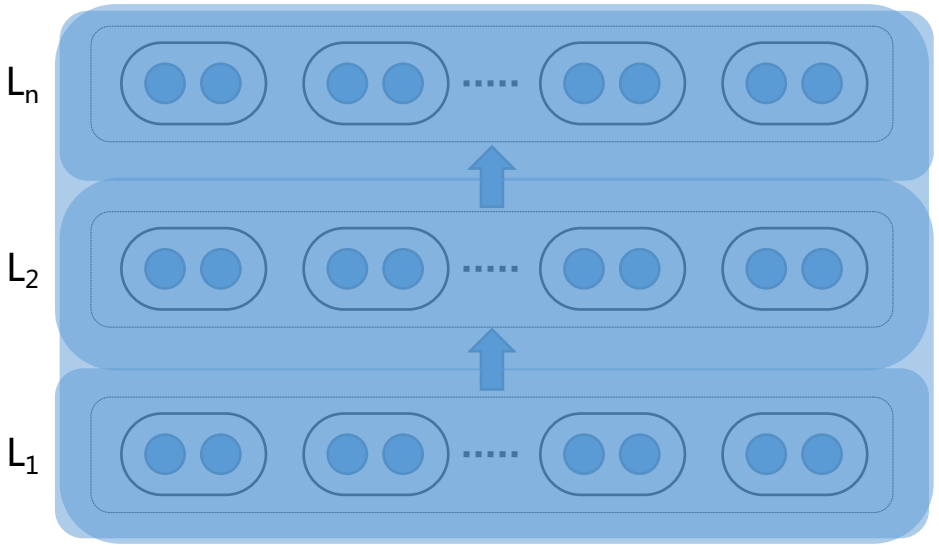
更多精调方法

- 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)



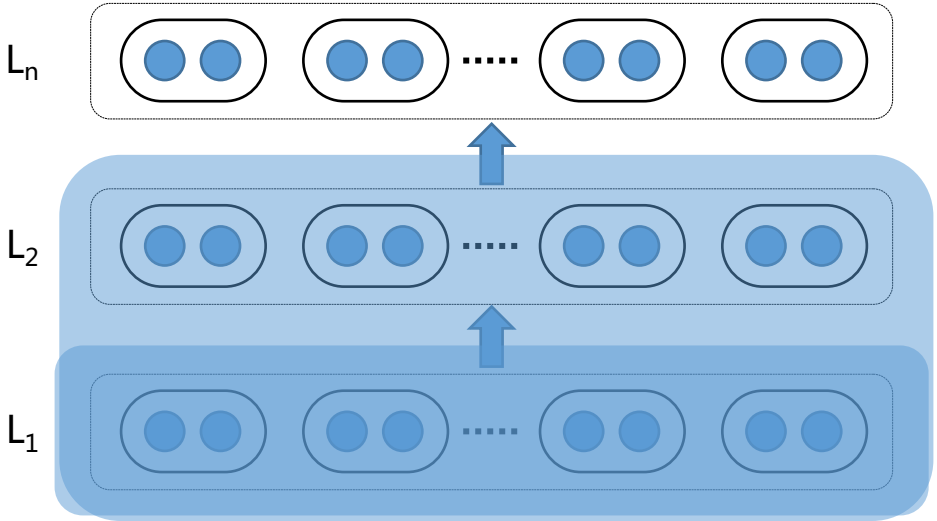
更多精调方法

- 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)

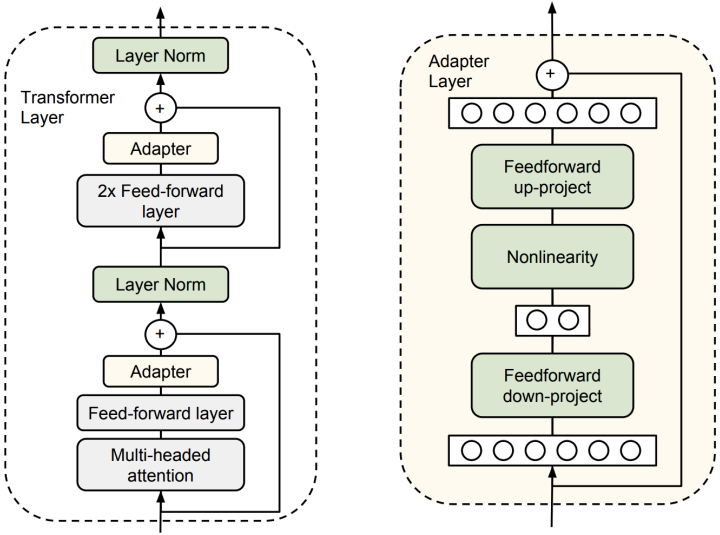


更多精调方法

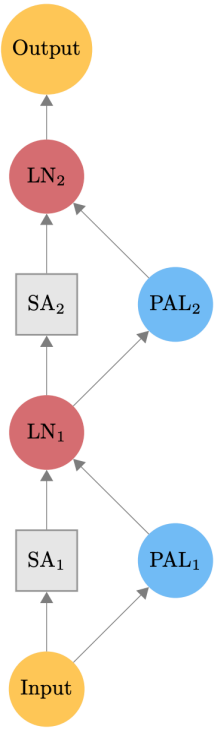
- 目标：既要适应目标任务，又要避免重写预训练模型
- 方法
  - 只精调最后一层，固定其它层 (Long et al., ICML 2015)
  - 每次只精调一层，固定其它层 (Felbo et al., EMNLP 2017)
  - 自顶向下逐层解冻 (Howard and Ruder, ACL 2018)
    - 其它策略
      - 学习率预热
      - 二次预训练：在目标领域未标注数据上精调语言模型
  - 将目标模型每层的参数和激活与预训练模型进行比较，作为额外损失 (Wiese et al., CoNLL 2017)

更多精调方法

- 在Transformer中增加适配器（Adapter）



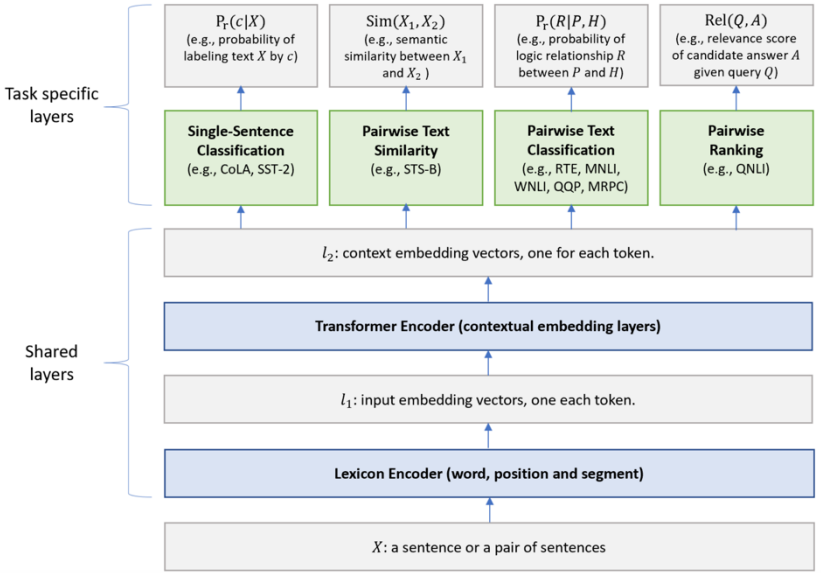
(Houlsby et al., ICML 2019)



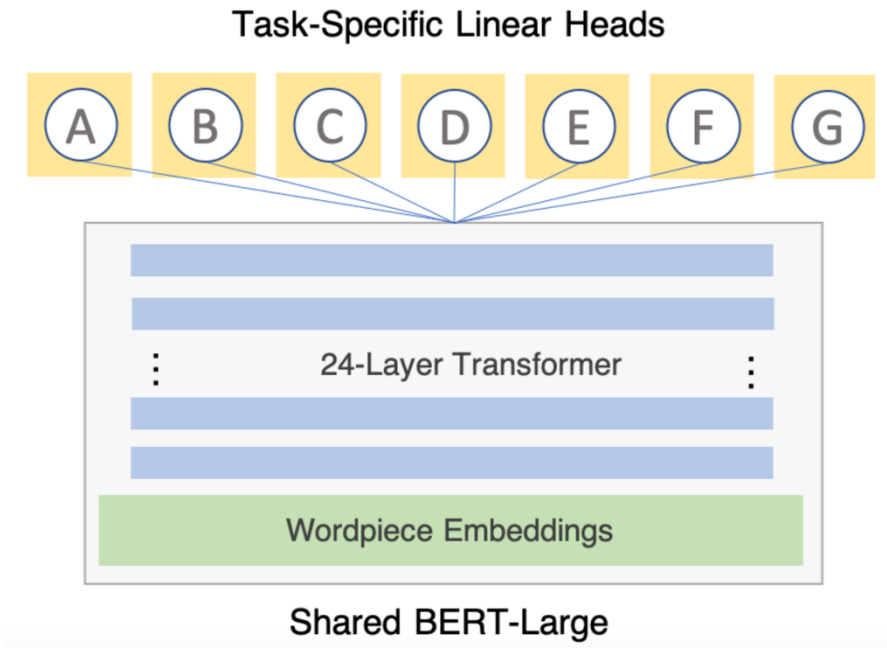
(Stickland and Murray, ICML 2019)

多任务学习

- 使用多任务学习框架，综合利用多种类型数据



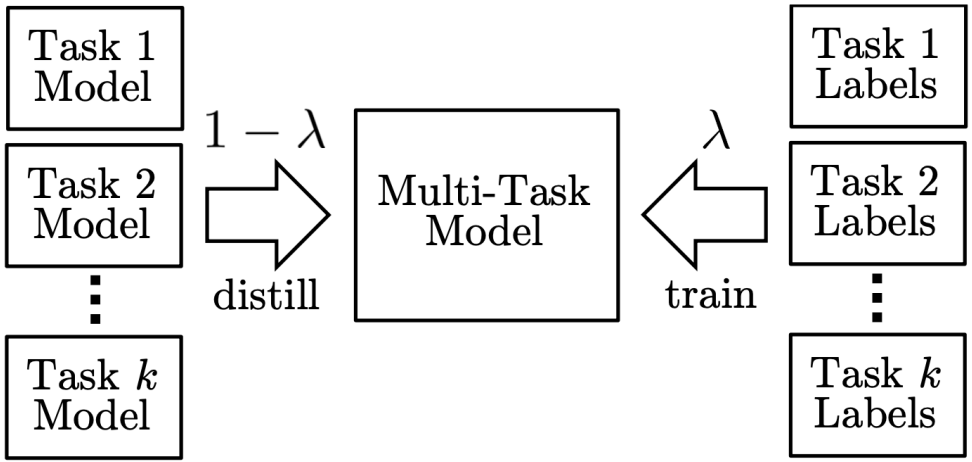
(Liu et al., ACL 2019)



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

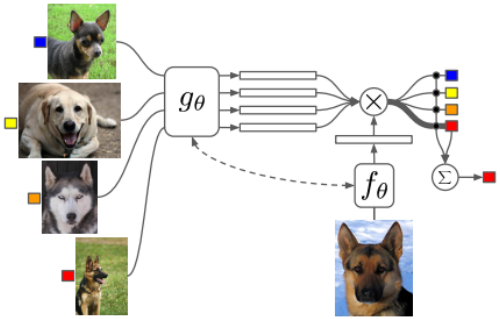
多任务学习

- 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



小样本学习

- Few-Shot Sequence Labeling with Label Dependency Transfer and Pair-wise Embedding (Hou et al., arXiv:1906.08711)
  - 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



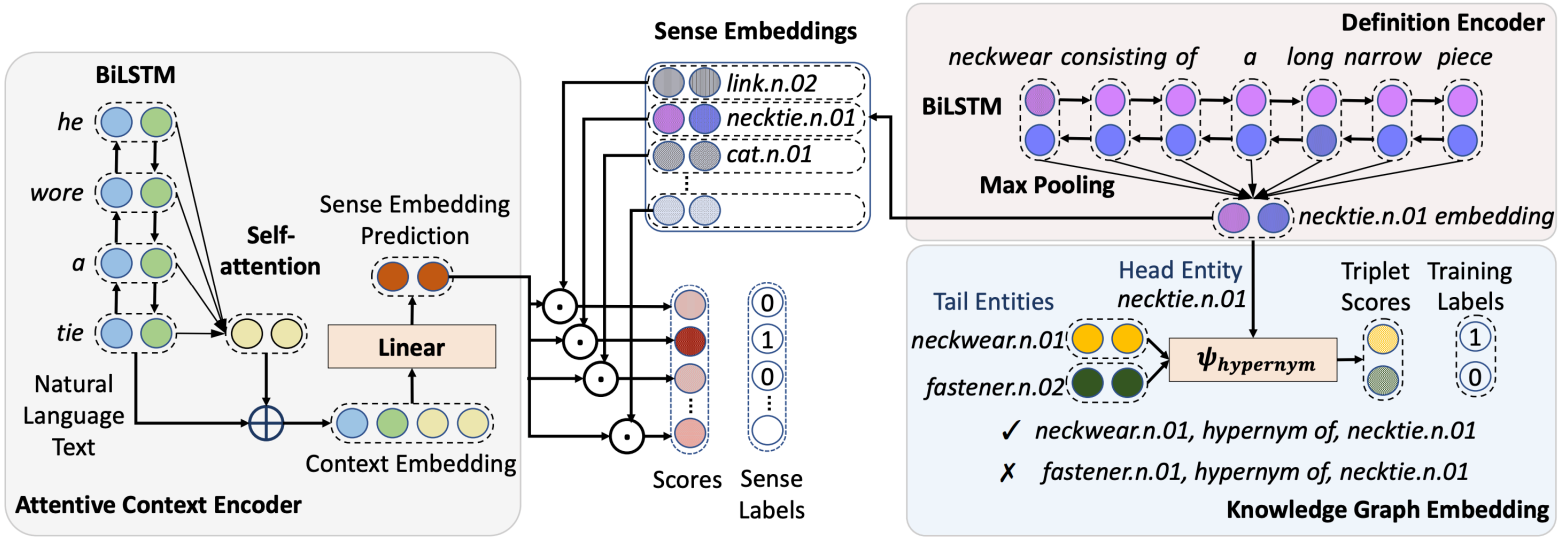
x: Will it snow$_{[B\text{-}weather]}$ the$_{[B\text{-}time]}$ day$_{[I\text{-}time]}$ after$_{[I\text{-}time]}$ tomorrow$_{[I\text{-}time]}$?

Trans: O → O → B-weather → B-time → I-time → I-time → I-time

0.38 0.07 0.10 0.53 0.41 0.41

零样本学习

- 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



大纲

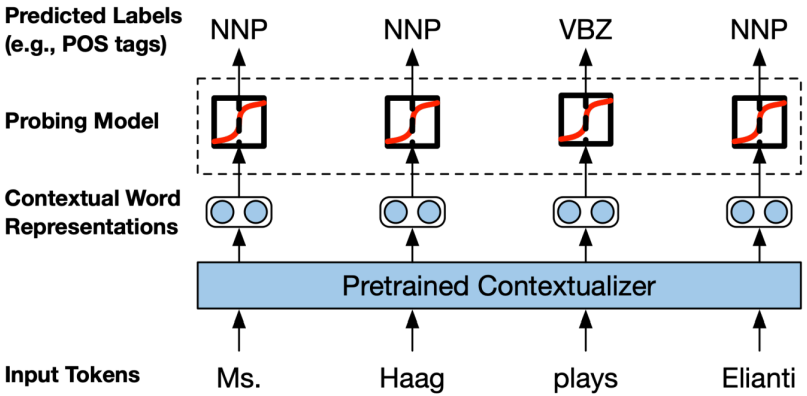
- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 预训练模型的应用
- 预训练模型的分析
- 预训练模型的挑战

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

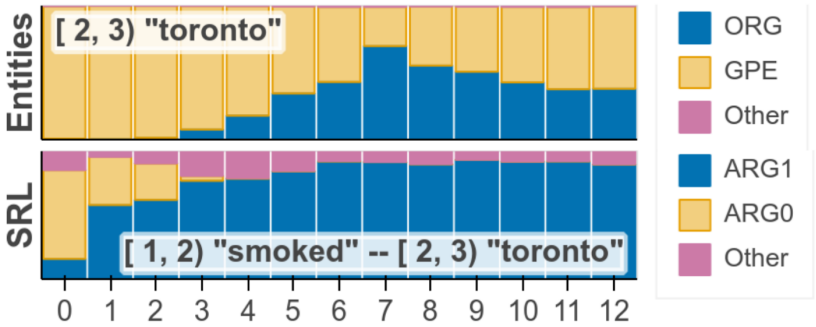
- Linguistic Knowledge and Transferability of Contextual Representations (Liu et al., NAACL 2019)
  - 在16个下游任务中进行实验
    - 固定预训练模型，作为特征提取器
    - 最上层只使用任务相关的线性分类器
  - 结论
    - 预训练模型在大部分任务中表现优异
    - 除了需要细粒度语言知识的任务
      - 如语法检查、NER、并列成分识别等
    - RNN模型（如ELMOs）的上层和任务相关
    - Transformer表现并非如此
    - 在相关有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

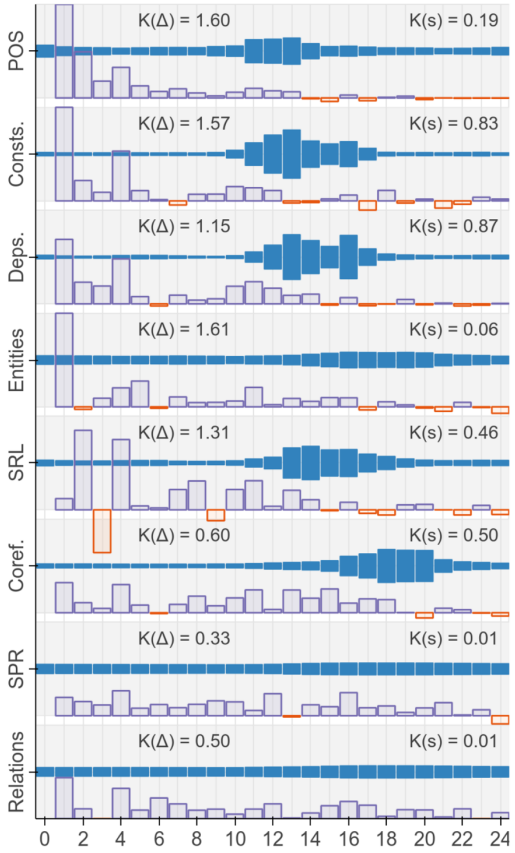


级联任务探针

- 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

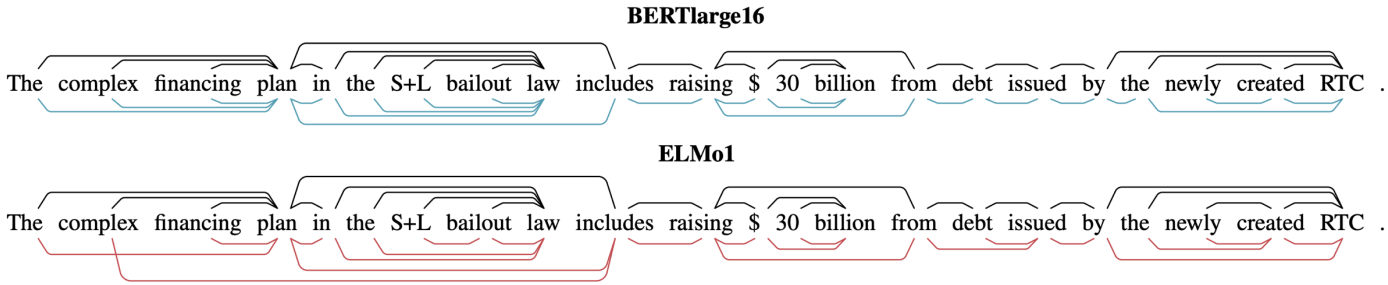
(a) he smoked **toronto** in the playoffs with six hits, seven walks and eight stolen bases ...





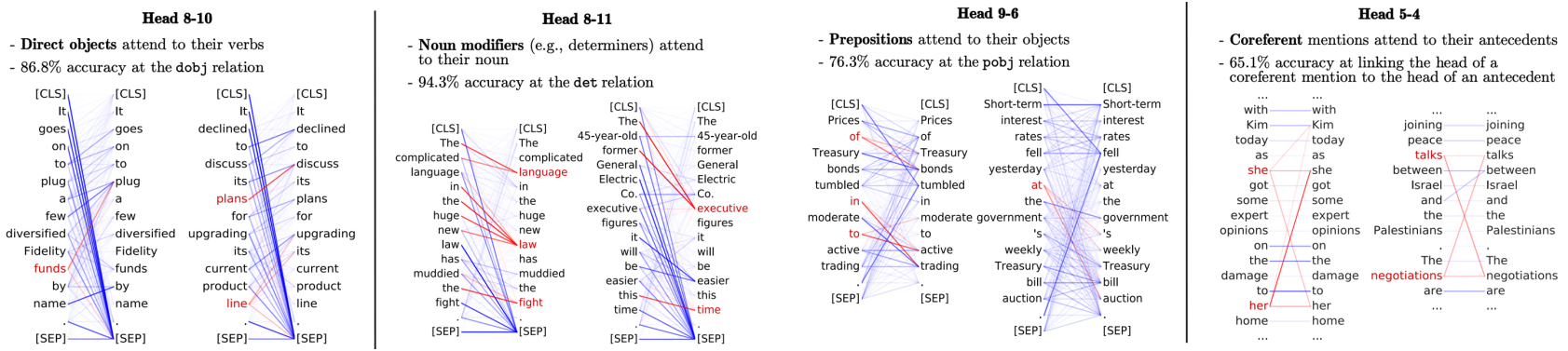
上下文词向量探针

- 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



注意力探针

- What does BERT look at? An Analysis of BERT's Attention (Clark et al., arXiv:1906.04341)
  - 某些Head中，注意力同样蕴含了句法，甚至指代的信息

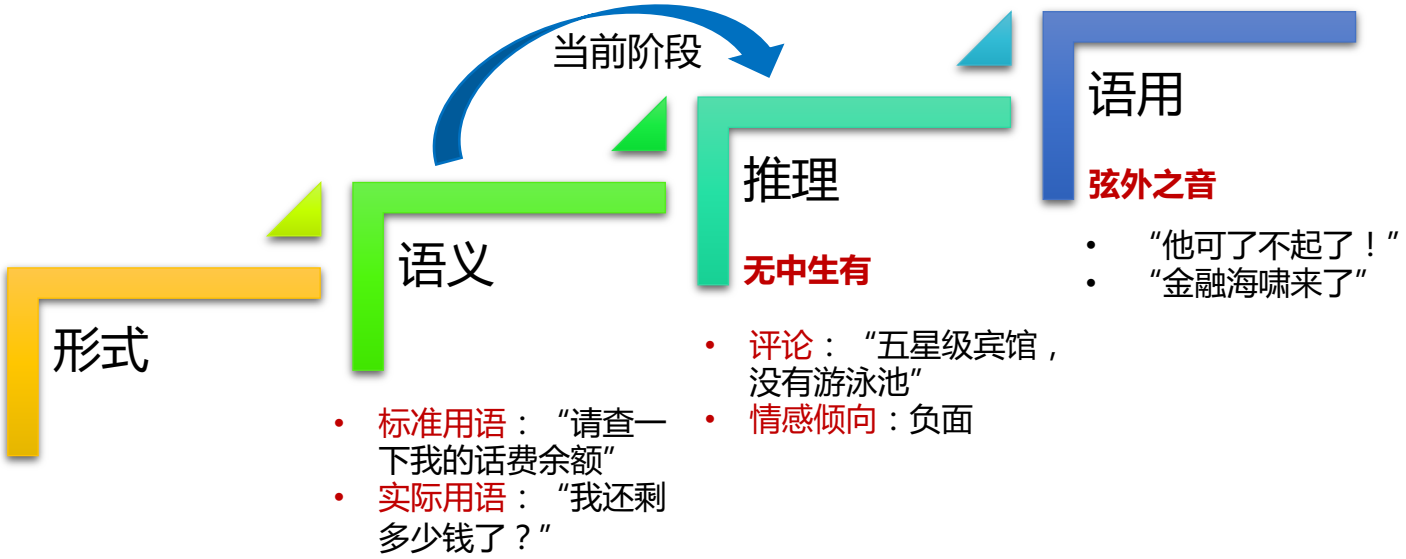


大纲

- 传统词向量预训练
- 上下文相关词向量
- NLP中的预训练模型
- 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

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？
- 如何进行模型压缩与加速？
  - DistilBERT效果不佳
  - ALBERT推理速度无优势
- 如何在seq2seq任务中使用BERT？
  - 多遍采样 (Wang and Cho, NeuralGen 2019)、重排序
- 如何对长文档进行表示？
- 如何应对对抗攻击？

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

相关资源

- NAACL 2019 Tutorial: Transfer Learning in Natural Language Processing
  - https://github.com/huggingface/naacl_transfer_learning_tutorial
- 清华NLP组Pre-trained Languge Model (PLM) 论文汇总
  - https://github.com/thunlp/PLMpapers
- HuggingFace开源Transformers（PyTorch BERT→PyTorch Transformers → Transformers）
  - https://github.com/huggingface/transformers
- 哈工大讯飞联合实验室发布的中文BERT
  - https://github.com/ymcui/Chinese-BERT-wwm

谢谢！

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

参考文献

- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model. JMLR 2003.
- David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. JMLR 2003.
- Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Lai. Class-based N-gram Models of Natural Language. CL 1992.
- Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, Ting Liu. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. CoNLL Shared Task 2018.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, Christopher D. Manning. What Does BERT Look At? An Analysis of BERT's Attention. arXiv:1906.04341.
- Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, Quoc V. Le. BAM! Born-Again Multi-Task Networks for Natural Language Understanding. ACL 2019.
- Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel P Kuksa. Natural Language Processing (Almost) from Scratch. JMLR 2011.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. Cross-Lingual Machine Reading Comprehension. EMNLP 2019.

参考文献

- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc V Le, Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. ACL 2019.
- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis. JASIST 1990.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann. Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. EMNLP 2017.
- Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu. Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources. COLING 2014.
- Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, Ting Liu. Cross-lingual Dependency Parsing Based on Distributed Representations. ACL 2015.
- Braden Hancock, Clara McCreery, Ines Chami, Vincent S. Chen, Sen Wu, Jared Dunnmon, Paroma Varma, Max Lam, and Chris Ré. Massive Multi-Task Learning with Snorkel MeTaL: Bringing More Supervision to Bear.

参考文献

- John Hewitt, Christopher D Manning. A Structural Probe for Finding Syntax in Word Representations. NAACL 2019.
- Yutai Hou, Zhihan Zhou, Yijia Liu, Ning Wang, Wanxiang Che, Han Liu, Ting Liu. Few-Shot Sequence Labeling with Label Dependency Transfer and Pair-wise Embedding. arXiv:1906.08711.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. ICML 2019.
- Jeremy Howard, Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. ACL 2018.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding. arXiv:1909.10351.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy. SpanBERT: Improving Pre-training by Representing and Predicting Spans. arXiv:1907.10529.
- Sawan Kumar, Sharmistha Jat, Karan Saxena, Partha Pratim Talukdar. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings. ACL 2019.
- Guillaume Lample, Alexis Conneau. Cross-lingual Language Model Pretraining. arXiv:1901.07291.

参考文献

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao. Multi-Task Deep Neural Networks for Natural Language Understanding. ACL 2019.
- Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, Noah A Smith. Linguistic Knowledge and Transferability of Contextual Representations. NAACL 2019.
- Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, Guoping Hu. Generating and Exploiting Large-scale Pseudo Training Data for Zero Pronoun Resolution. ACL 2017.
- Yijia Liu, Wanxiang Che, Yuxuan Wang, Bo Zheng, Bing Qin, Ting Liu. Deep Contextualized Word Embeddings for Universal Dependency Parsing. TALLIP 2019.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Michael Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, Ping Wang. K-BERT: Enabling Language Representation with Knowledge Graph. arXiv:1909.07606.

参考文献

- Mingsheng Long, Yue Cao, Jianmin Wang, Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. ICML 2015.
- Zhimao Lu, Haifeng Wang, Jianmin Yao, Ting Liu, Sheng Li. An Equivalent Pseudoword Solution to Chinese Word Sense Disambiguation. ACL 2006.
- Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Ming Zhou. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. EMNLP 2019.
- Bryan McCann, James Bradbury, Caiming Xiong, Richard Socher. Learned in Translation: Contextualized Word Vectors. arXiv:1708.00107.
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. ICLR 2013.
- Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocký, Sanjeev Khudanpur. Recurrent neural network based language model. INTERSPEECH 2010.
- Andriy Mnih, Geoffrey E Hinton. A Scalable Hierarchical Distributed Language Model. NIPS 2008.
- Jeffrey Pennington, Richard Socher, Christopher Manning. Glove: Global Vectors for Word Representation. EMNLP 2014.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher G Clark, Kenton Lee, Luke Zettlemoyer. Deep Contextualized Word Representations. NAACL 2018.

参考文献

- Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, Noah A. Smith. Knowledge Enhanced Contextual Word Representations. EMNLP 2019.
- Matthew E Peters, Sebastian Ruder, Noah A Smith. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. arXiv:1903.05987.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018.
- Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. NeurIPS Workshop 2019.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tieyan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. arXiv:1905.02450.
- Asa Cooper Stickland, Iain Murray. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. ICML 2019.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, Jifeng Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. arXiv:1908.08530.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. ICCV 2019.

参考文献

- Siqi Sun, Yu Cheng, Zhe Gan, Jingjing Liu. Patient Knowledge Distillation for BERT Model Compression. arXiv:1908.09355.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration. arXiv:1904.09223.
- Ian Tenney, Dipanjan Das, Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. arXiv:1905.05950.
- Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, Ting Liu. Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing. EMNLP 2019.
- Alex Wang, Kyunghyun Cho. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. NeuralGen 2019.
- Georg Wiese, Dirk Weissenborn, Mariana Neves. Neural Domain Adaptation for Biomedical Question Answering. CoNLL 2017.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G Carbonell, Ruslan Salakhutdinov, Quoc V Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. ERNIE: Enhanced Language Representation with Informative Entities. ACL 2019.
- Sanqiang Zhao, Raghav Gupta, Yang Song, Denny Zhou. Extreme Language Model Compression with Optimal Subwords and Shared Projections. arXiv:1909.11687