从深度匹配到深度问答

复旦大学 张奇

目录

从PC到移动

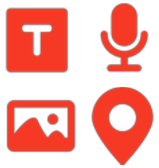
设备

手机、Pad

任何可联网的设备

（眼镜、车载、音箱、服务机器人）

输入



关键字、语音
图像、POI

自然语言(语音)、环境

信息载体

网页、APP

信息、知识、常识

关键词 **10条链接**

搜索（用户阅读）

问句 **答案**

问答（机器阅读）

无处不在的问答引擎

Q: 北京处理违章需要带什么证件？

Q：汽车没电了打不着火怎么办？
A：首先需要一根跨接电线，然后将两车的车头面对面...

Q：大红袍是什么茶？
A：乌龙茶。

Q：学生证买火车票一年能用几次？
A：4次。

来自搜索的问题分布

用户需求强，难度大

①约1/4用户查询是问答需求
②更长尾，更偏向自然语言描述

查询类别	搜索请求占比
问答	24%
医疗	9%
小说	9%
教育	7%
人物	7%
知识	4.5%
购物	4%

猕猴桃和奇异果有什么区别	腰肌劳损挂什么科
眼镜蛇和眼镜王蛇的区别	蜂蛹泡酒的功效
人有多少颗牙齿	什么是幽门螺旋杆菌
没有房产证的酒店拆迁如何补偿	手机怎么注册淘宝账号
低烧可以进藏吗	落地签证是什么意思
影响发电量的三大因素	充电器可以托运吗
esp是什么	苹果手机信号不好怎么办
神经绷紧怎么回事	烫伤的水泡会自然消吗
顶的网络意思	网速突然变慢的原因
大米生虫子了还能吃吗	在家原地跑步能减肥吗
网很卡是什么原因	交首付款注意事项
折耳兔怎么养	故乡鲁迅发生的变化的原因
石家庄市银监局投诉电话是多少	疝气会影响生育吗

搜索问答的挑战

面临复杂的互联网文档，文档质量、文本长度千差万别，机器阅读理解难度更大，更具挑战性。

任务	阅读对象	文档质量	文本长度	文档是否一定存在答案
Squad1.0/2.0	Wiki段落	高质量	平均长度约138个词，200词以下占比86%，300词以下占比98%	1.0一定存在答案 2.0不一定存在答案
搜索问答	网页文本	类型多样，质量差异大，大量UGC内容质量低	平均正文约460字（除视频等非文本页） 超过600字的占比约1 /3	不一定存在答案 单个搜索结果中的答案可能错误

关键问题1： 从全网信息中精准找到包含答案的结果 → 更高的相关性要求 → 深度匹配

关键问题2： 从多样的网页中精准抽取出答案 → 更复杂的阅读理解 → 深度问答

深度匹配

深度匹配 – 数据来源

搜狗问问

地理学 地球科学

为什么地球上没能同时进化出两种以上的智慧生命?

匿名用户 6万 次浏览 | 2018.08.09 15:10 提问

我来回答

最佳答案

由提问者、达人共同推荐

企鹅问答

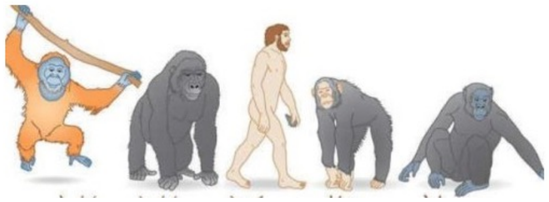
企业

2018.08.09 15:35 回答

首先，在地球进化45+亿年的绵延历史中，有人类特征的动物只是最近400-600万年的事情，人类其实是地球本身演化和地球生物进化的集大成者。

如果把地球历史比作长江，从唐古拉山脉的源头到入海口处，约6280公里，人类历史出现在距离入海口的8000多米处。

在这之前，人类前身的动物的祖先经历了从单细胞、多细胞、简单动物、脊椎动物、哺乳动物到灵长类和猿……，人类从其他猿类区分开的时候已经是大约400-600万年前的事情了，人类的基因里已经凝聚了天地万物之灵！起点很高！



iPhone

问：ios12.0.1系统为什么升级后信号变差?

如题 昨天升级后明显信号变得很差 许多APP点开都不能刷新

发布日期 2018年10月13日 上午1:25

回复 我也有此问题 (1)

知乎

化学 化工 化学元素 元素周期表 化学专业

集齐元素周期表所有元素，大概需要多少钱?

突然开的脑洞，不需要剂量特别多，想收藏全套。

谢谢大家提醒，考虑到半衰期，稀有度和放射

关注问题 写回答 邀请回答 44 条评论 分享 举报 ···

查看全部 293 个回答

氧化气氖氚

27,008 人赞同了该回答

嗯…一群巨佬已经把非放射性元素收的差不多了……我就大约上两张图，稍微估个价吧。

先在最前面说一嘴，价格是要可以以此流通的，所以很多半衰期很短的元素不在我的讨论中（完全没发流通对不对…）。

会做到评论90%回复，剩下10%有的可能不知道回点啥……有点也是没来及回复。私信可以100%回复，欢迎提问。

哇哇哇，一晚上赞数15个诶，好激动，在最后补一个铯融化的视频。
以后说不好还会上稀有气体的辉光视频。

开心，不到一天快百赞了，第一次回答很开心，慢慢发稀有气体的辉光视频，有审核，所以可能有些慢。（视频已更完）

深度匹配

Apple 官方支持：

如果您的 iPhone、 iPad 或 iPod touch 开不了机或死机怎么办

用户：

iPhone 死机了怎么办

苹果8突然卡死，主界面不能动，在线急等

我的**肾7**突然无法操作，这种情况怎么解决？

用户输入与标准问法差别很大

深度匹配

小孩子发烧**38**度怎么办 ←→ 脑袋测得出的东西叫智商，脑袋测不出的东西叫智慧；么办

北京大学 ←→ 北京**的**

盐酸**氯**丙嗪 ←→ 盐酸**氯**

定金 ←→ **订**金

眼睛看得到的地方叫视线，眼睛看不到的地方叫视野；

耳朵听得到的动静是声音，耳朵听不到的动静是声誉；

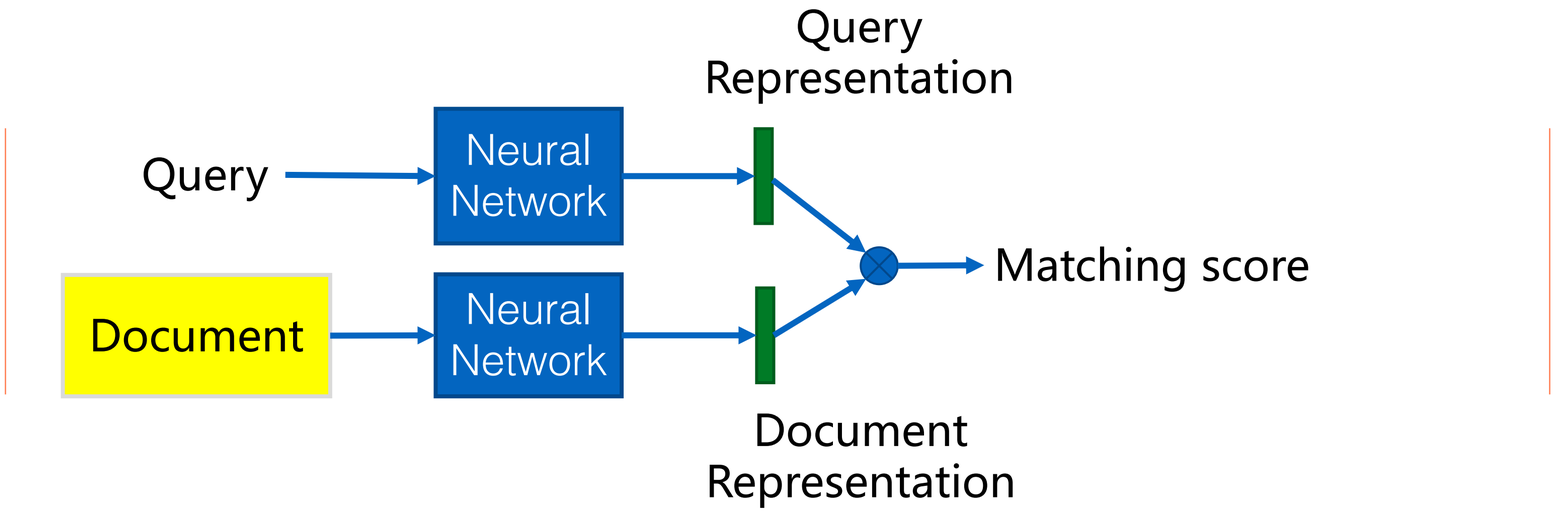
嘴里说得出来的话叫内容，嘴里说不出来的话叫内涵；

……

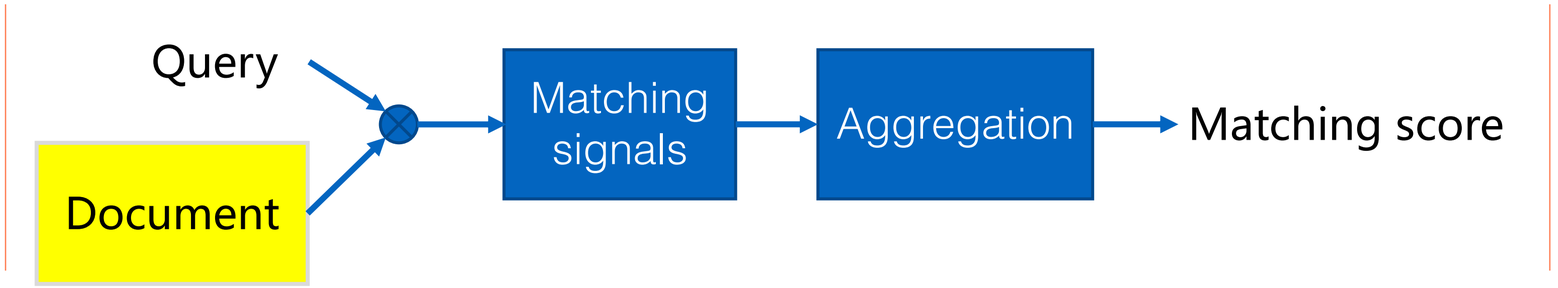
失之毫厘，差以千里!

深度匹配

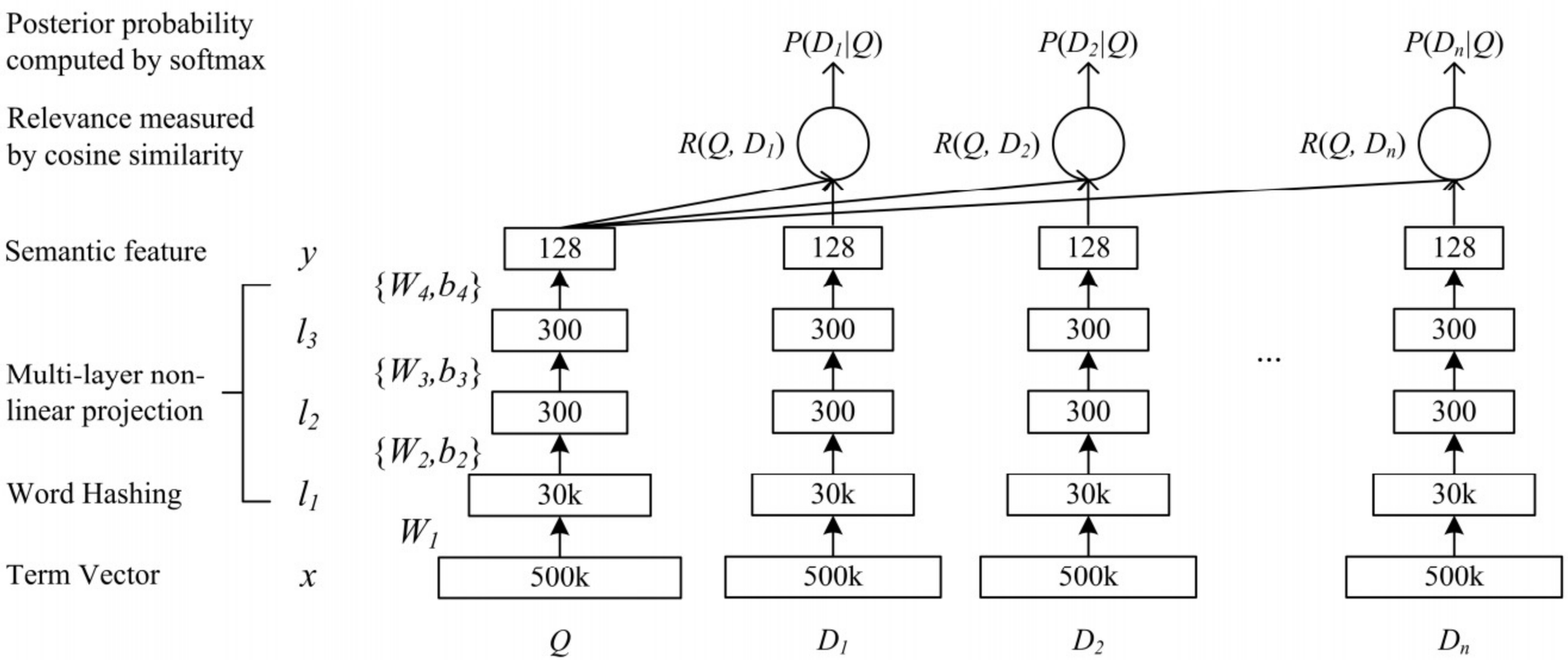
基于句子表示的方法



基于交互关系的方法

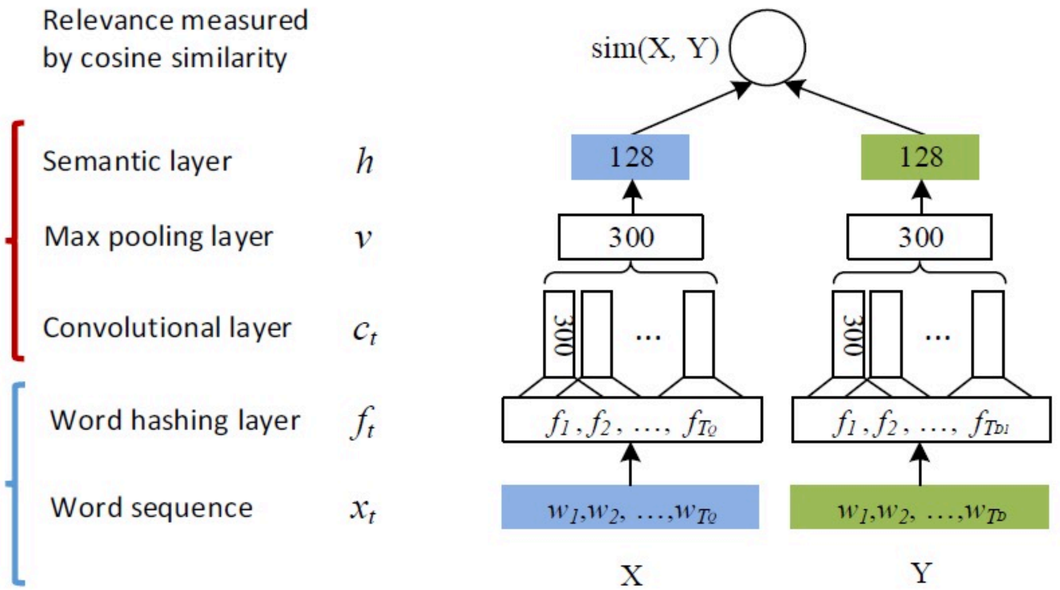


基于句子表示的方法

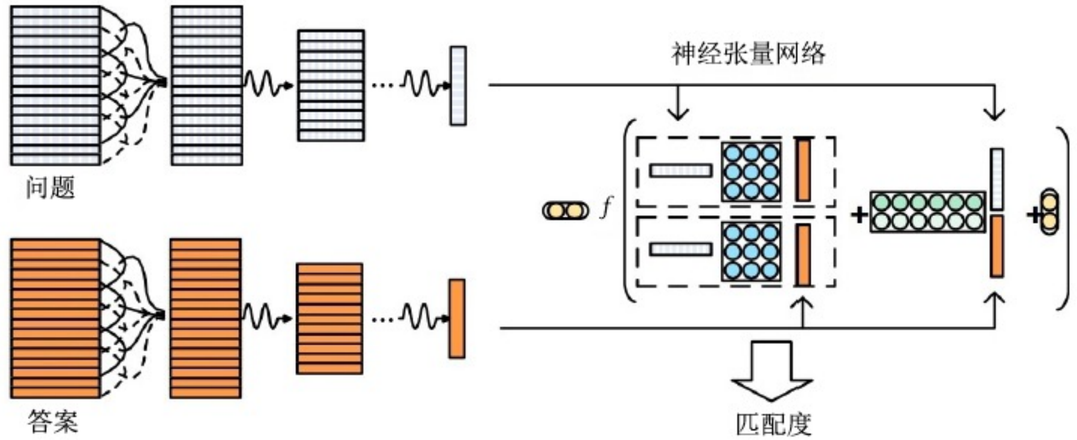


DSSM: Learning Deep Structured Semantic Models for Web Search using Click-through Data (Huang et al., CIKM'13)

基于句子表示的方法

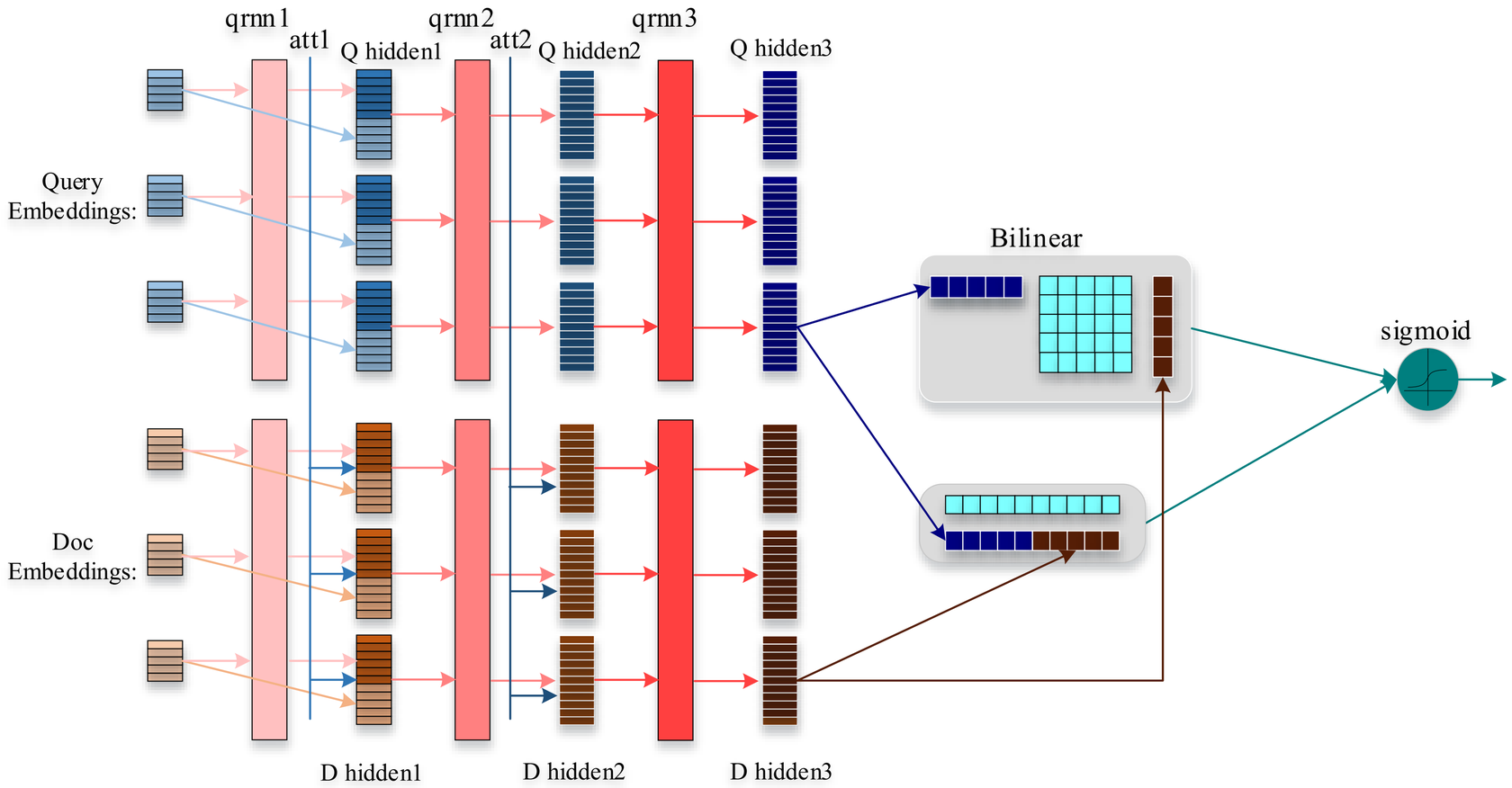


CNN-DSSM



CNTN

基于句子表示的方法

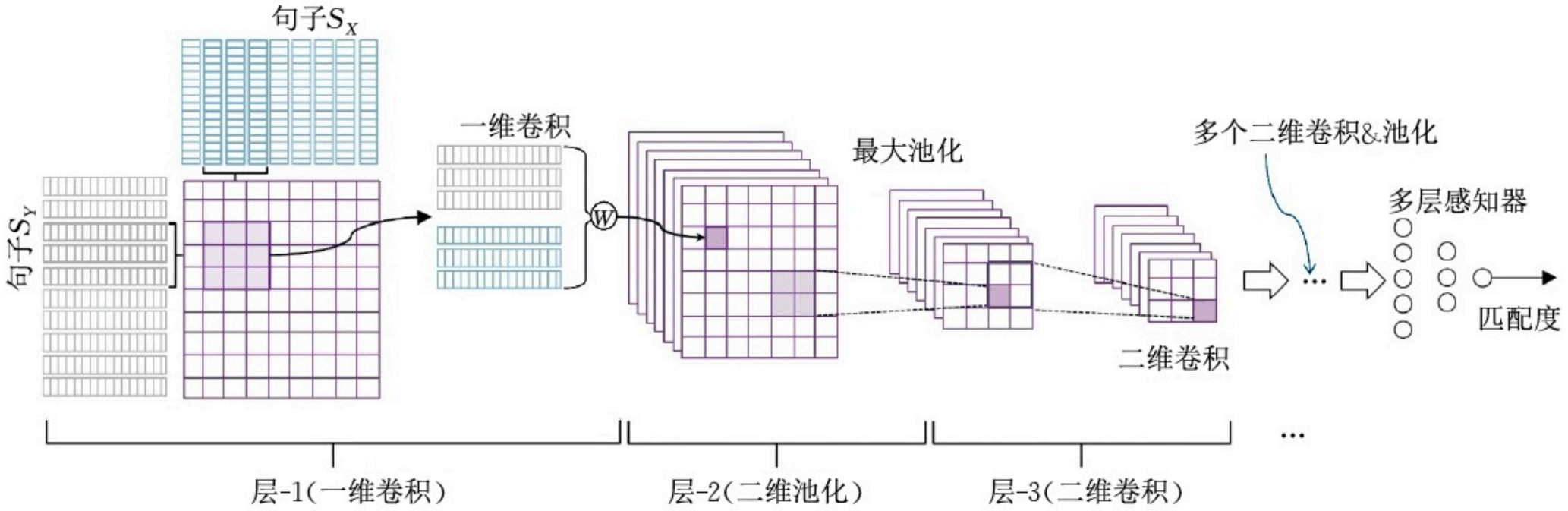


Query和Doc的多层交互，三层QRNN提供了由浅到深的语义编码能力

Attention机制建模Query和Doc的语义一致性，利用Query重构Doc上下文表示的能力，语义信息层层递进

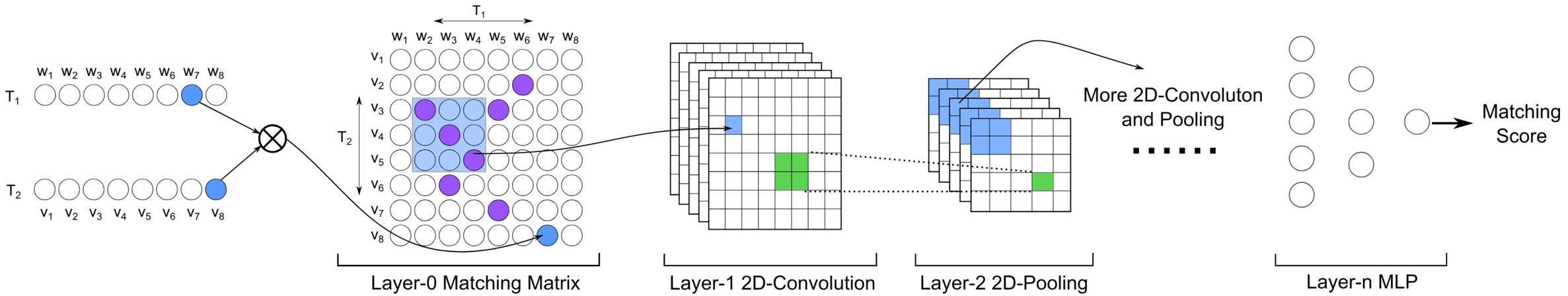
多层表示学习 QRNN-ATT

基于交互关系的方法



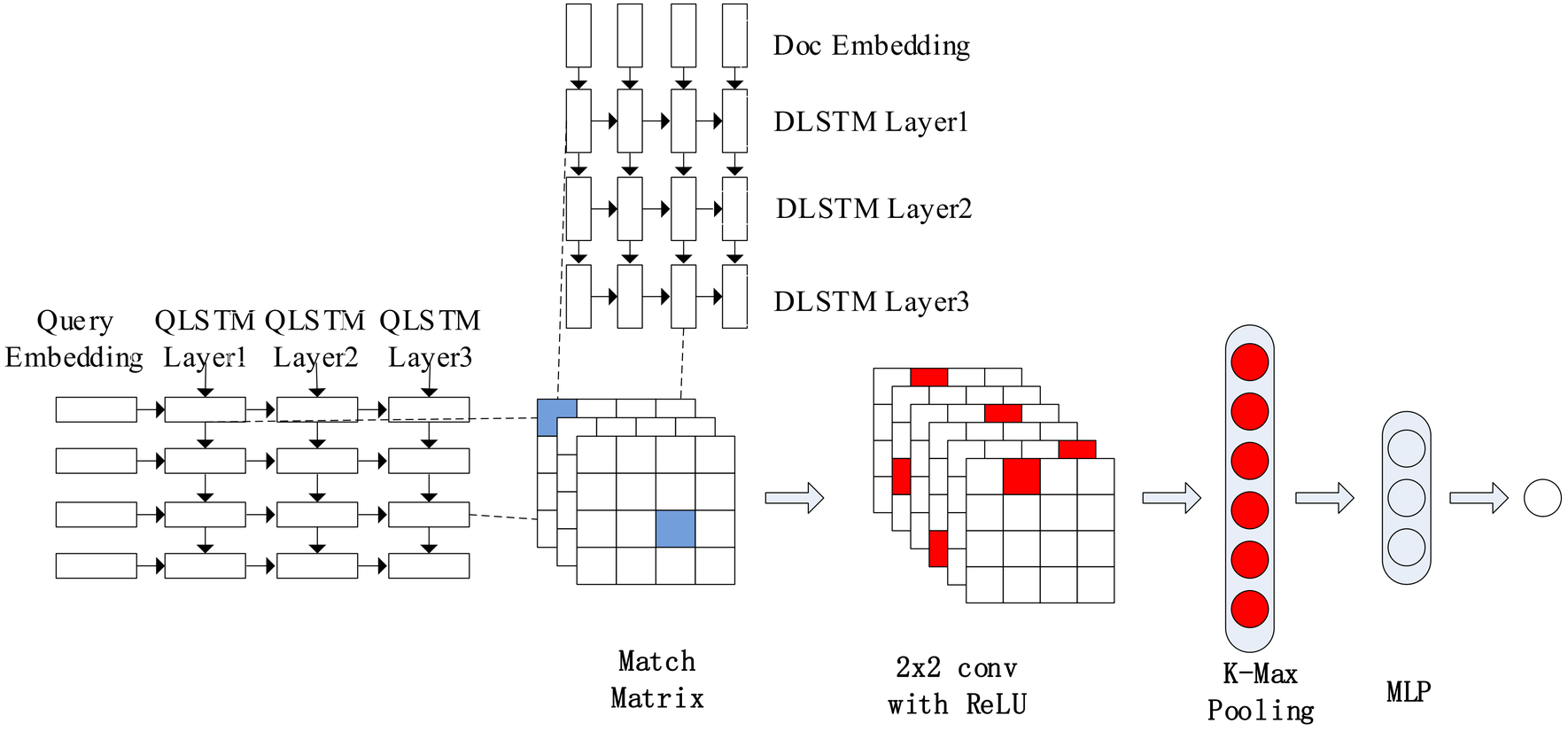
ARC-II

基于交互关系的方法



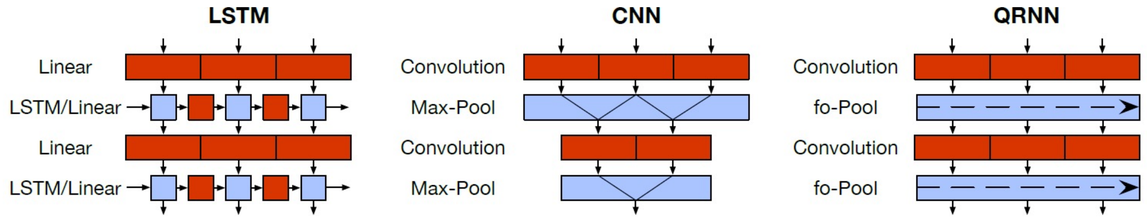
MatchPyramid

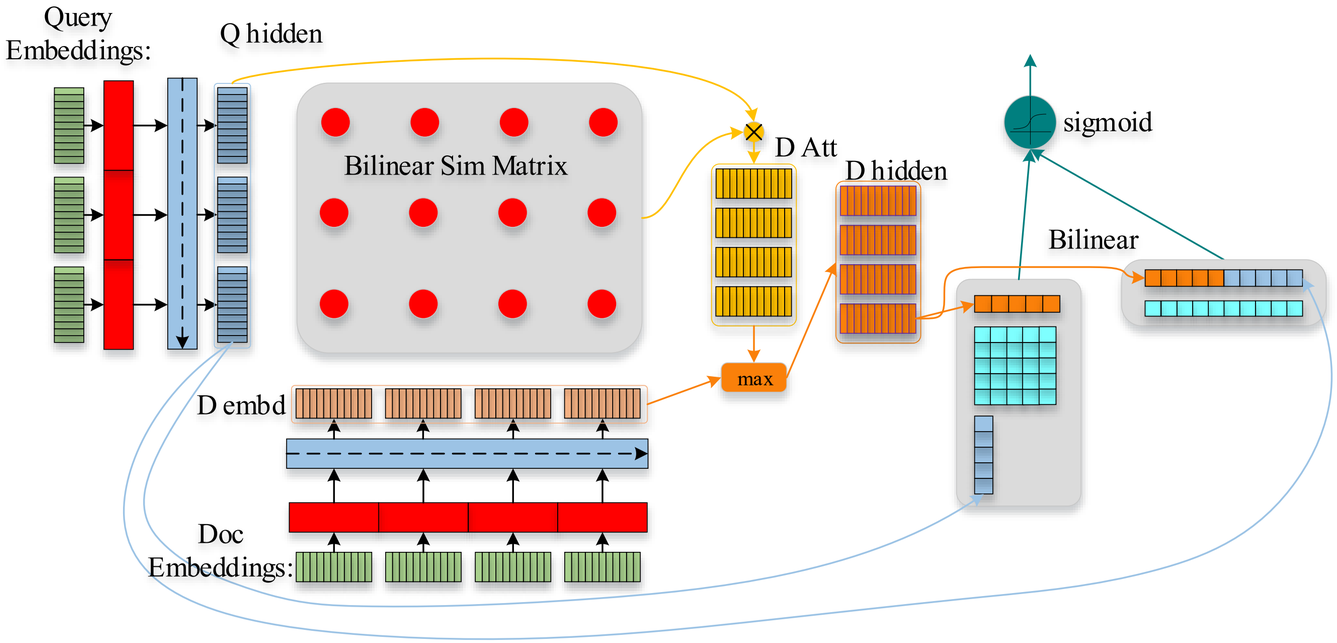
基于交互关系的方法



多层匹配学习 MV-LSTM

基于交互关系的方法





基于QRNN和注意力机制的表示学习

序列表示，3倍加速于LSTM

注意力机制加强Query-Doc的交互

单层QRNN-ATT

公开数据集上的表现

深度匹配模型与传统BM25特征差距不大

- 数据深度不足，Robust 04 包含60万词，50万篇文档。

基于交互关系的方法优于基于表示的方法

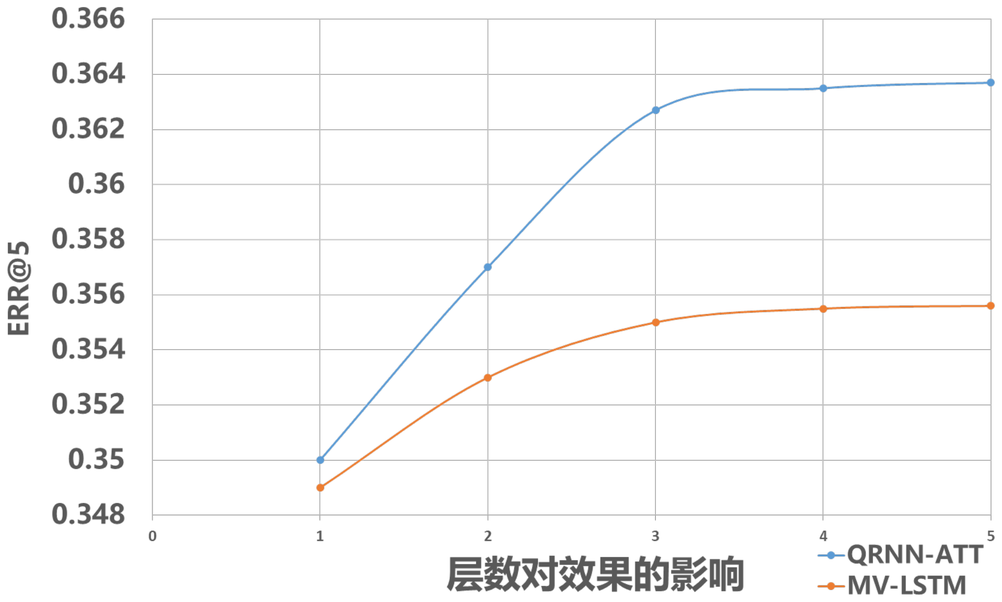
- 基于交互关系的方法引入了词精确命中，模糊命中，模板匹配等强信息，缓解了训练数据不足的问题。

		Method	MAP	nDCG@20
Traditional IR		BM25	0.255	0.418
Deep Learning Approaches to Matching	Representation Learning	DSSM	0.095	0.201
		CDSSM	0.067	0.146
		ARC-1	0.041	0.066
		MV-LSTM	0.119	0.185
	Matching Function Learning	ARC-2	0.067	0.147
		MatchPyramid	0.189	0.330
		Match-SRNN	0.203	0.374

Robust 04

(庞亮 深度文本匹配综述 计算机学报 2016)

方法比较-多层模型



		Method	ERR@1	ERR@5
Traditional IR		BM25	0.181	0.331
Deep Learning Approaches to Matching	Representation Learning	单层QRNN-ATT	0.198	0.350
		三层QRNN-ATT	**0.208**	**0.363**
	Matching Function Learning	单层MV-LSTM	0.196	0.349
		三层MV-LSTM	0.202	0.355

Sogou longtail dataset

多层模型表示学习优于匹配学习，随层数增加，表示学习效果提升更明显

- Representation Learning模型能够更好地设计多层间交互，提升模型拟合能力

机器阅读理解

机器阅读理解

周杰伦 编辑词条 添加义项 | 同义词 | 收藏 | 分享

周杰伦，1979年1月18日出生于台湾省新北市，中国台湾流行乐男歌手、音乐人、演员、导演、编剧、监制、商人。2000年被吴宗宪发掘并推出首张专辑《Jay》。2005年涉足影视，以电影《头文字D》获第42届台湾电影金马奖及第25届香港电影金像奖"最佳新人"奖。2007年成立JVR有限公司；同年，自编自导自演电影《不能说的秘密》，获得台湾电影金马奖年度台湾杰出电影奖。2008年凭借歌曲《青花瓷》获得第19届金曲奖最佳作曲人奖。2011年凭借专

2013年2月3日，周杰伦蜡像亮相上海杜莎夫人蜡像馆，成为首位拿到杜莎夫人蜡像馆亚洲地区所有分馆入住权的唯一华人男歌手（上海，香港，曼谷）；5月17日，"魔天伦"世界巡回演唱会在上海奔驰文化中心启动；6月1日，周杰伦配音《十万个冷笑话》第9、10集太乙真人；7月11日，周杰伦信主后首部自导自演的新片《天台爱情》首映。

周杰伦蜡像什么时间亮相上海杜莎夫人蜡像馆？

文章 + 问题 → 答案

机器阅读理解

2015年之前

- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

2015年之后

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- Maluuba NewsQA
- MS MARCO
- NAVER DSTC6-T1

From: "Towards the Machine Comprehension of Text" by Danqi Chen, 2017.

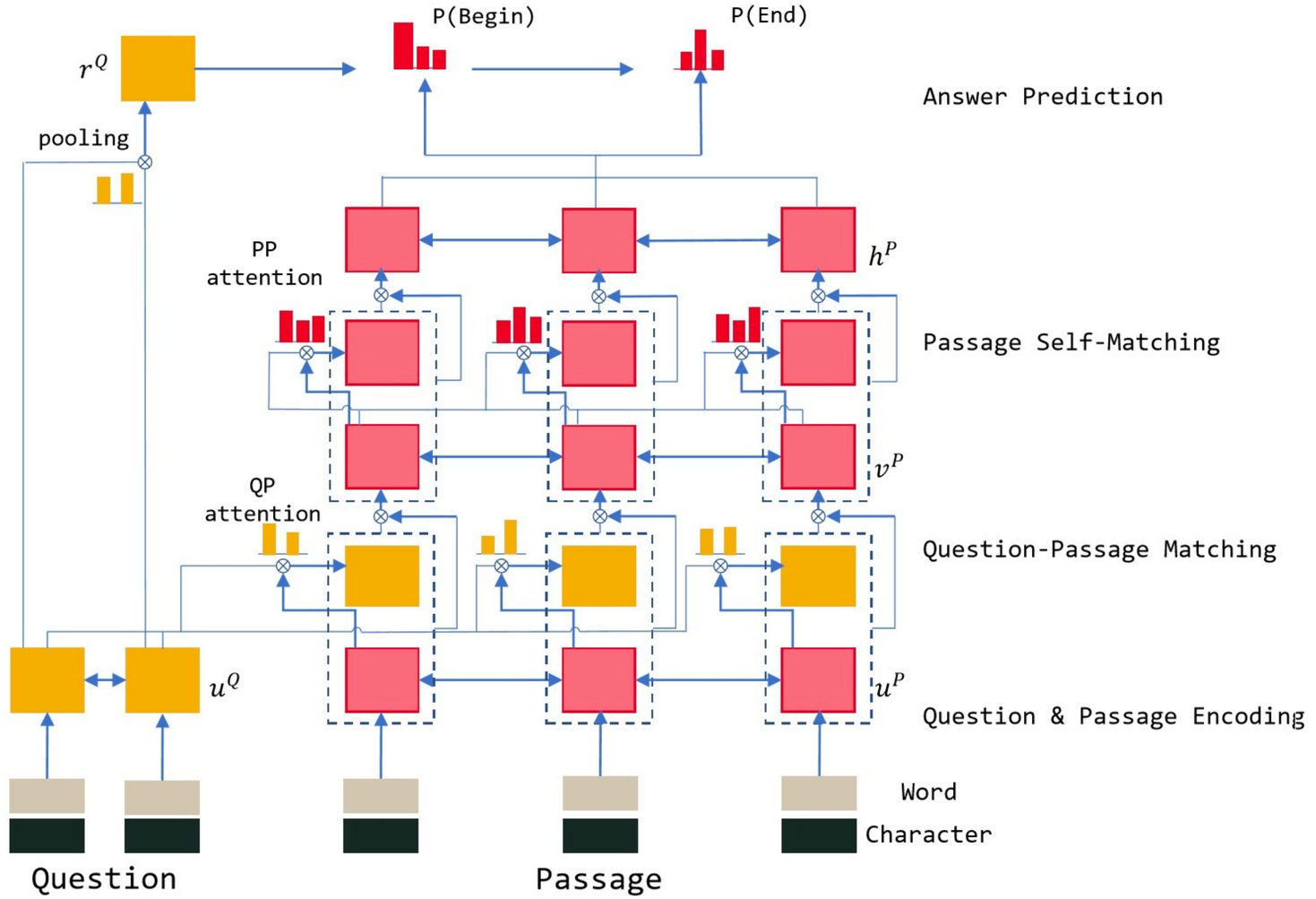
机器阅读理解

2015年之前的模型：单词匹配、逻辑回归、分类等

2015年之后的模型：**神经网络**

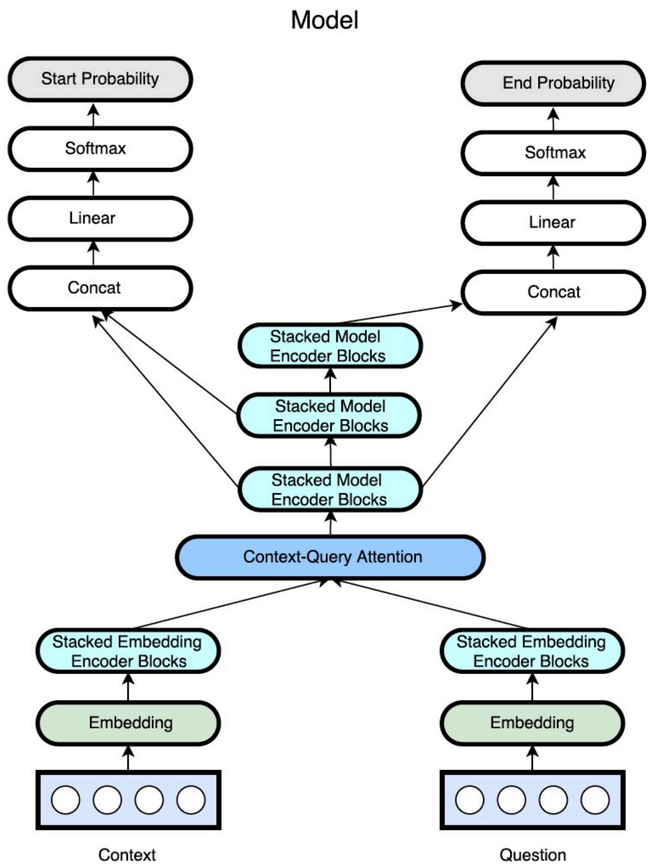
From: “Towards the Machine Comprehension of Text” by Danqi Chen, 2017.

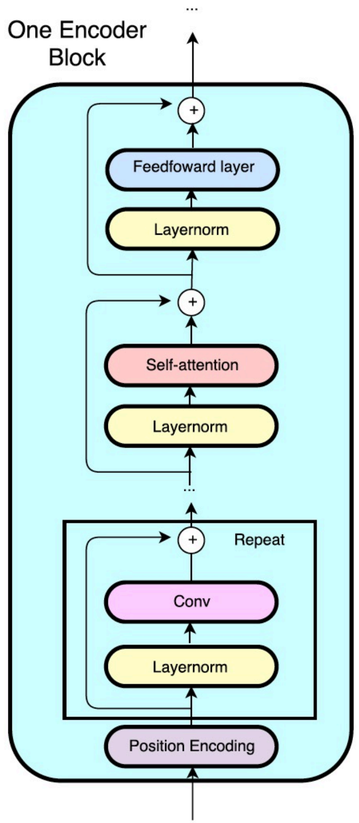
R-NET



R-NET: MACHINE READING COMPREHENSION WITH SELF-MATCHING NETWORKS, ACL 2017

Google QANET

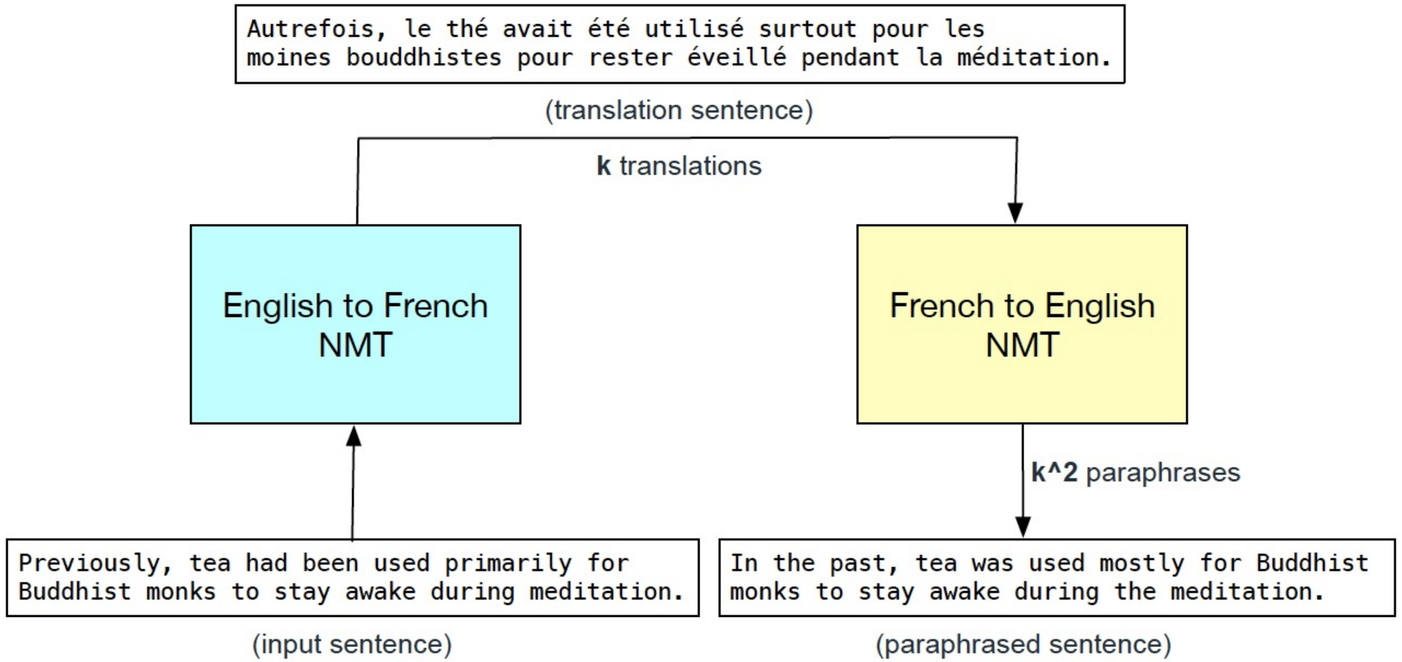




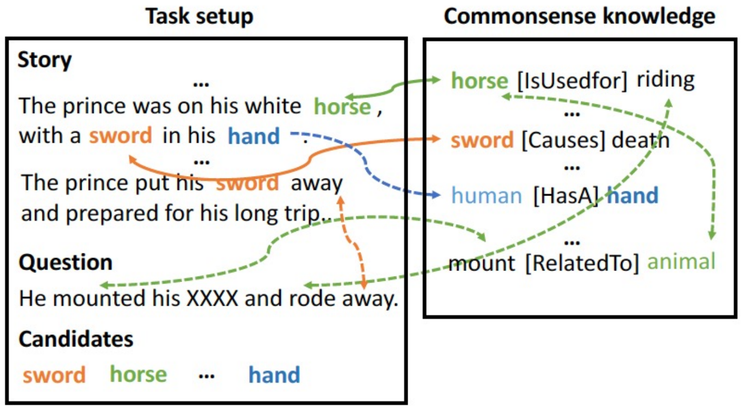
Yu, Adams Wei, et al. "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension." *arXiv preprint arXiv:1804.09541* (2018).

Google QANET

DATA AUGMENTATION BY BACKTRANSLATION



KnReader— Incorporating Commonsense Knowledge



Commonsense knowledge or factual background knowledge about entities and events

1. Knowledge Retrieval

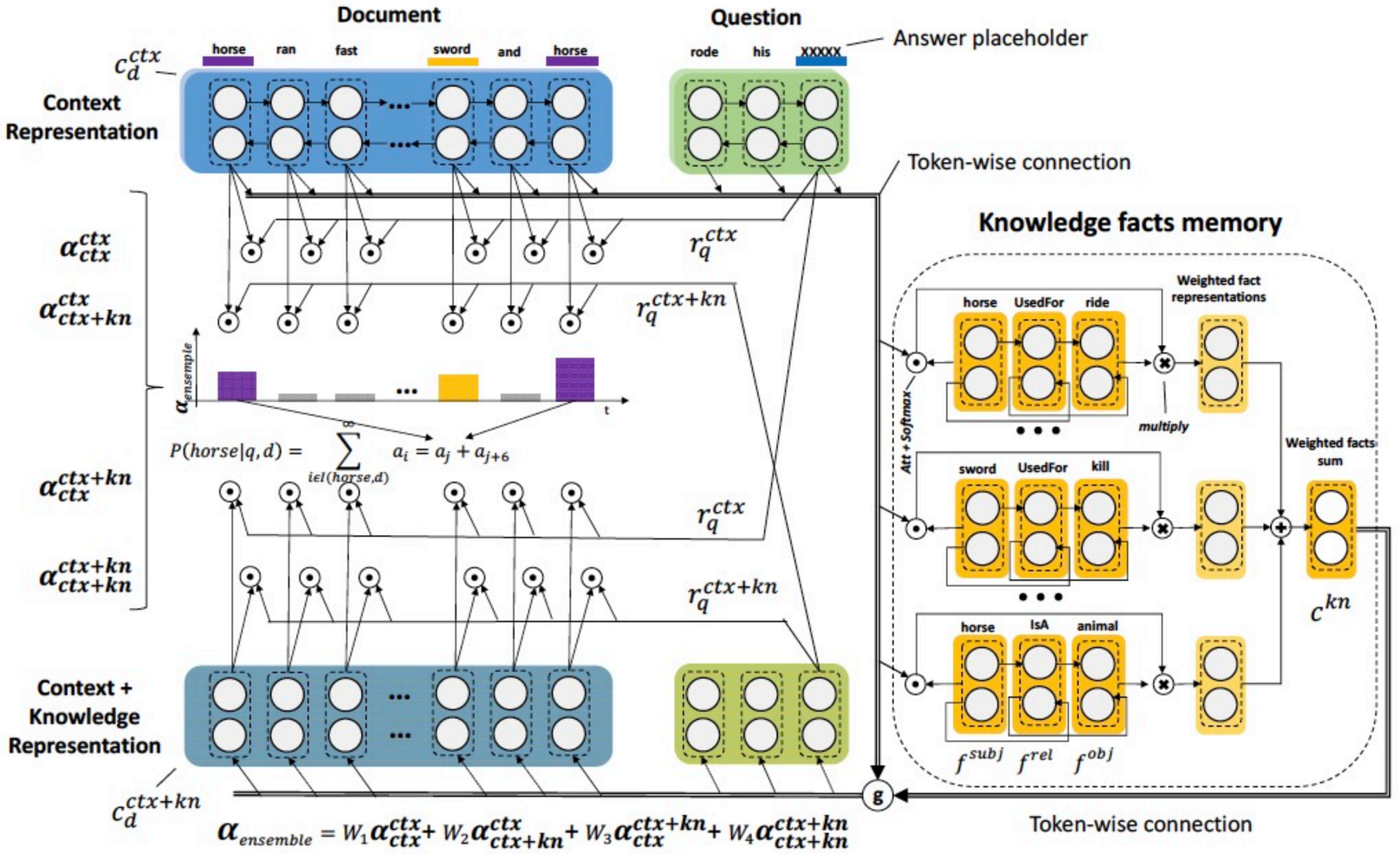
Performs fact retrieval and selects a number of facts

2. Knowledgeable Reader

Use Attention Sum Reader as one of the strongest core models for single-hop RC and extend it with a knowledge fact memory that is filled with pre-selected facts.

Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge, ACL 2018

KnReader— Incorporating Commonsense Knowledge



Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge, ACL 2018

事实类问答

面临的问题

①大量现有模型针对规范文本，假设文本中存在正确答案，而真实场景面临不存在答案的情况；

②现有模型很少考虑外部知识，对于涉及推理的问题，效果不好；

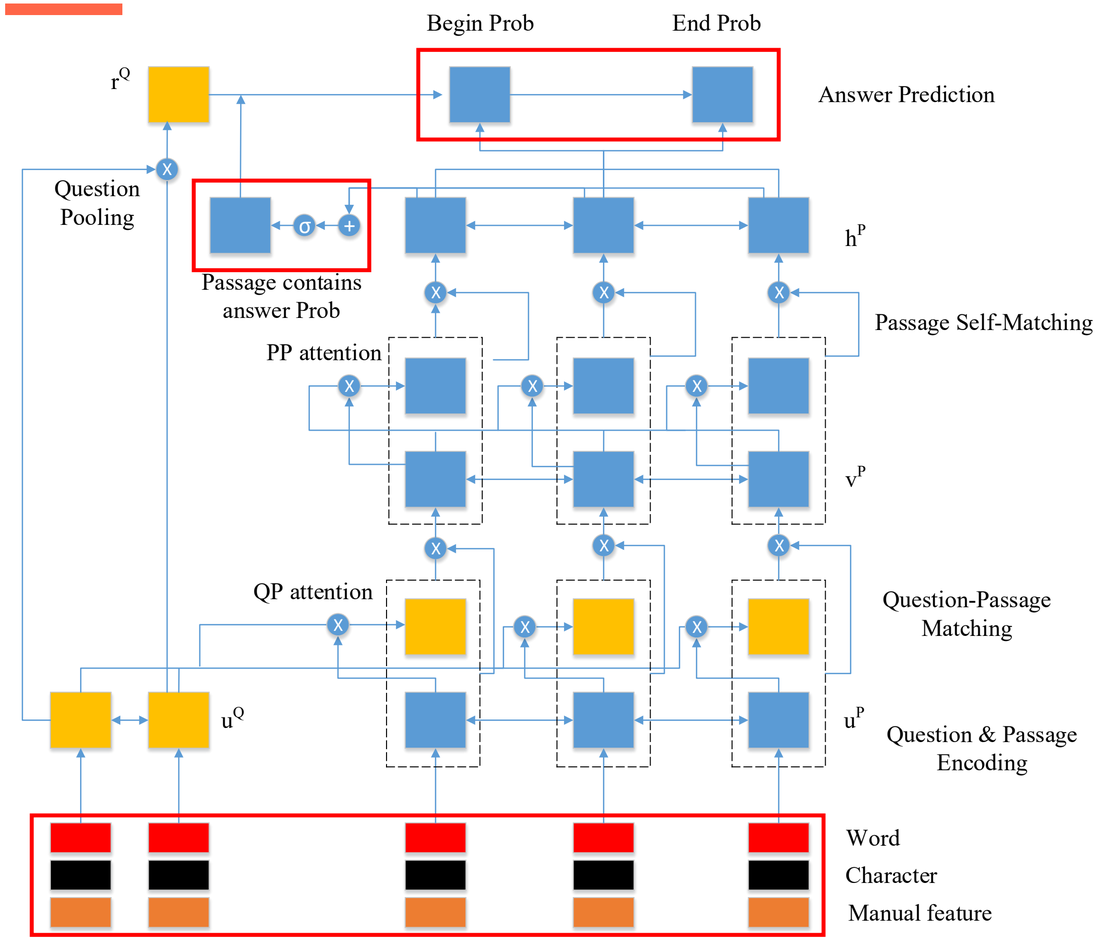
解决思路

①对答案存在性进行判断，避免错答；

②充分利用丰富的外部知识，不仅有利于处理复杂推理问题，也有助于答案存在性判断；

方案： 提出融合答案存在性判断、并结合外部知识的答案提取模型EK-RNet

EK-Rnet



1、融入passage包含答案的概率

2、在字、词表示基础上融入外部知识信息：全局 | 局部词频、是否为LAT、实体类型、数量词、来源网页质量等

EK-RNet模型，相比原始Rnet模型，在相同精度水平下（90%），召回提升30%

非事实类问答

面临的问题

① 网页通常包含长文本，现有阅读理解模型往往因长距离依赖导致丢失重要信息而提取错误答案

② 网页可能不存在正确答案，易过召

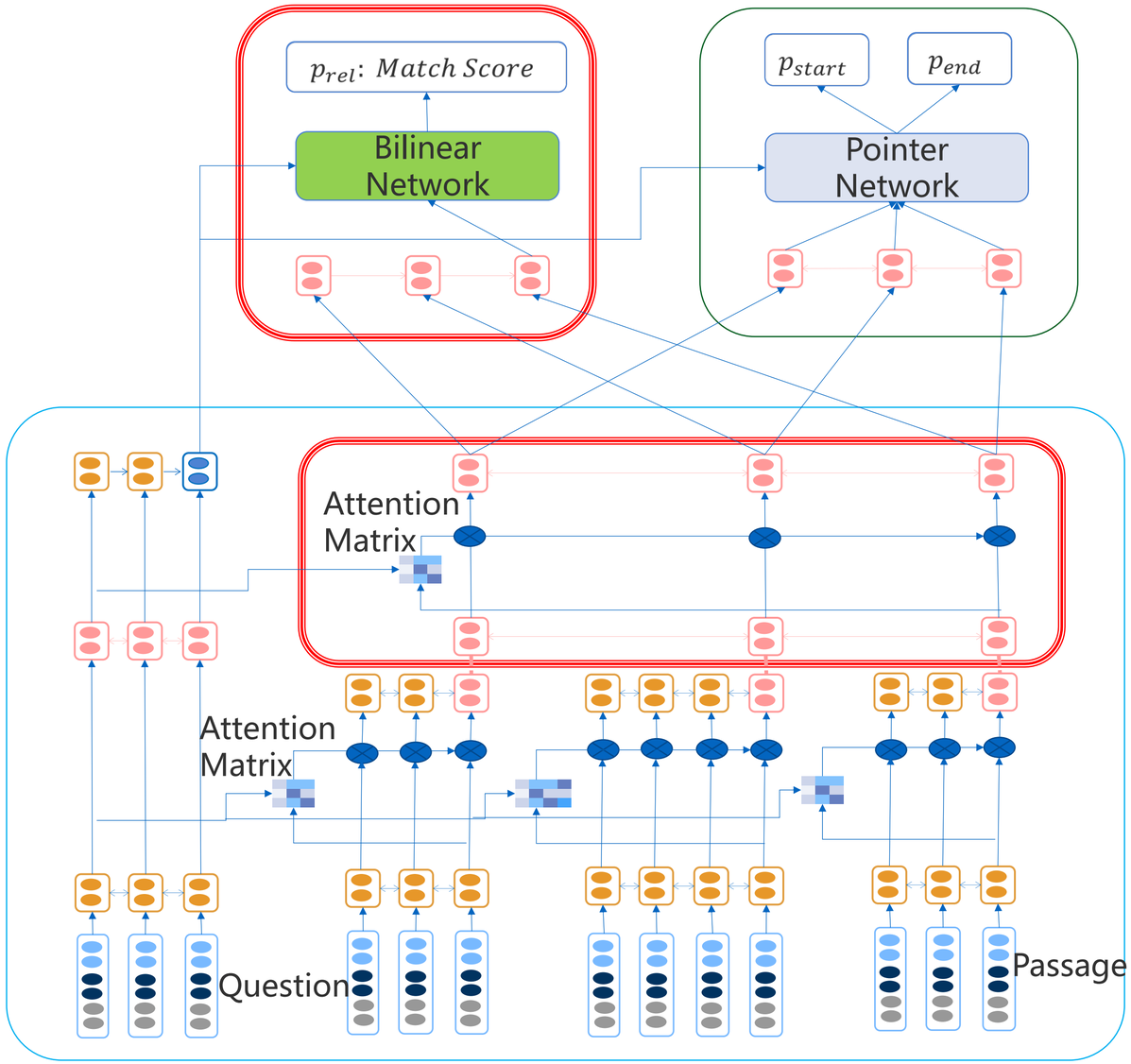
解决思路

① 针对长距离依赖和信息丢失问题，在以词为单位的表示和注意力机制基础上，融入以句子为单位的长距离信息传递机制，同时使用ELMO词嵌入，增加上下文相关表示能力

② 对答案存在性进行判断

方案：提出联合学习答案存在性判断与答案提取的分层匹配模型HM-LSTM

HM-LSTM



1.联合学习：
答案存在概率 + 答案起止位置概率

$$Loss = -y_{rel} log^{p_{rel}} - (1 - y_{rel}) log^{1-p_{rel}} - log^{p_{start}} - log^{p_{end}}$$

2.引入句子级别注意力机制
词注意力机制+句子注意力机制

3.使用上下文相关的词向量表示 ELMO

语言模型预训练

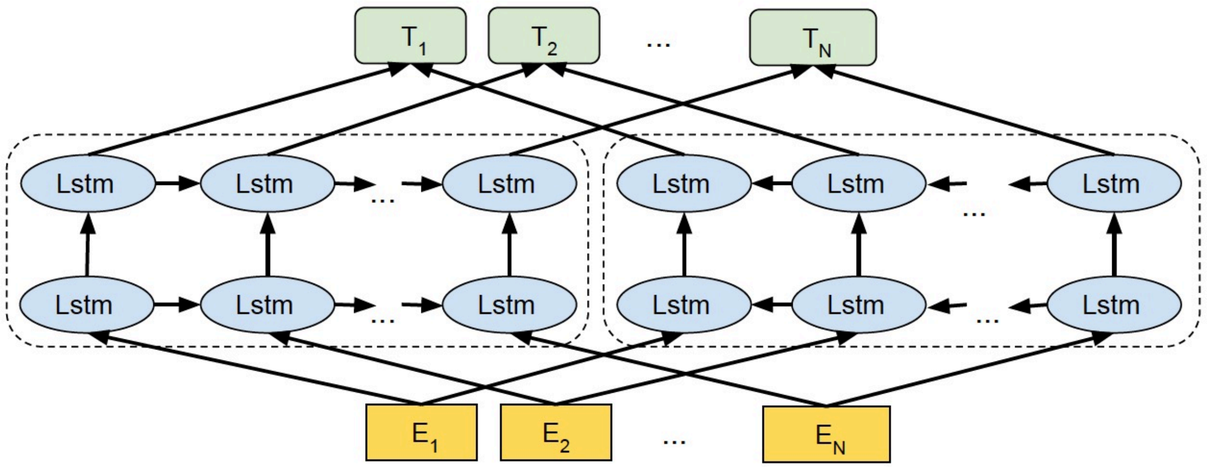
Word Embedding的缺陷

一词多义

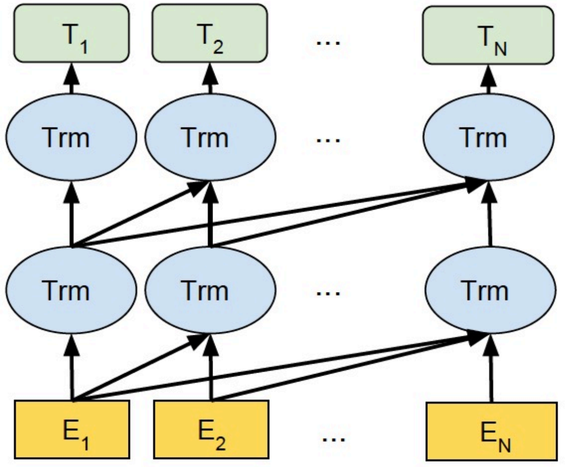
- 山上到处是盛开的杜鹃； 树林里传来了杜鹃的叫声。
- 把电视关上，我们要学习文件； 大家注意，下午要拿学习文件来。
- produce filets of smoked bass； exciting jazz bass player
- Jobs is the CEO of Apple; He finally ate the apple.

Word Sense Disambiguation

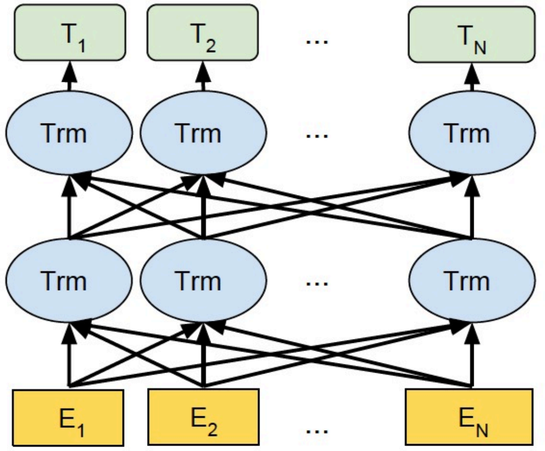
语言模型预训练方法



ELMO



OpenAI GPT



BERT

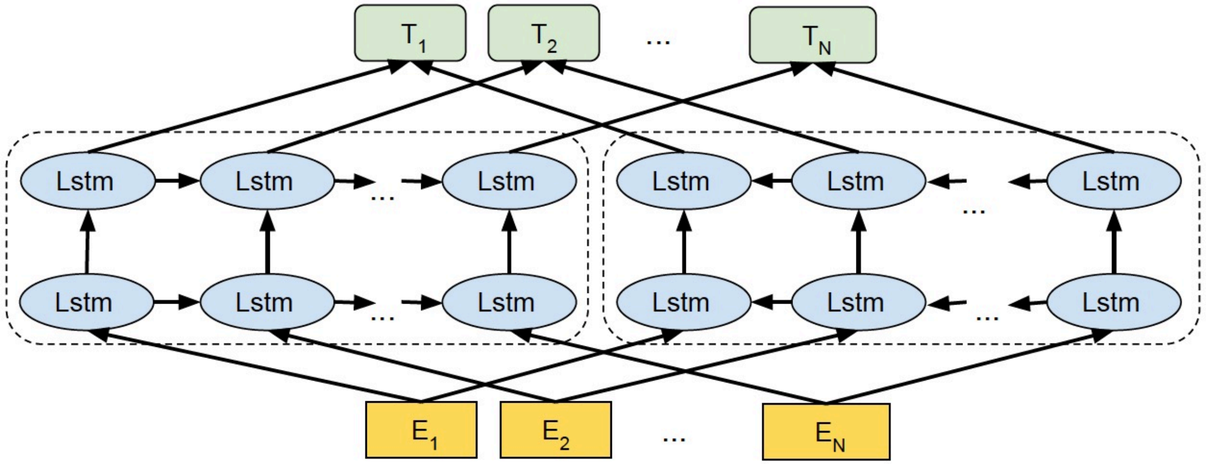
Embeddings from Language Models

- **传统Word Embeddings**
  - 每一个词只对应一个词向量

- **EMLo**
  - 利用预训练好的双向语言模型，然后根据具体输入从该语言模型中可以得到上下文依赖的当前词表
  - 对于不同上下文的同一个词的表示是不一样的

Peters, M. E. et al. Deep contextualized word representations. NAACL (2018).

Embeddings from Language Models



ELMO

双向的 LSTM 语言模型，一个前向和一个后向语言模型

目标函数: 取这两个方向语言模型的最大似然

训练时与任务无关

Peters, M. E. et al. Deep contextualized word representations. NAACL (2018).

Embeddings from Language Models

- 针对每个Token

$$
\begin{aligned}
R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\
&= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\},
\end{aligned}
$$

where $\mathbf{h}_{k,0}^{LM}$ is the token layer and $\mathbf{h}_{k,j}^{LM} = [\overrightarrow{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$, for each biLSTM layer.

- 针对某个特定的任务，将双向语言模型的每一中间层进行一个求和

$$
\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.
$$

- 有监督的 NLP 任务时，可以将 ELMo 直接当做特征拼接到具体任务模型的词向量输入或者是模型的最高层表示上。

Embeddings from Language Models

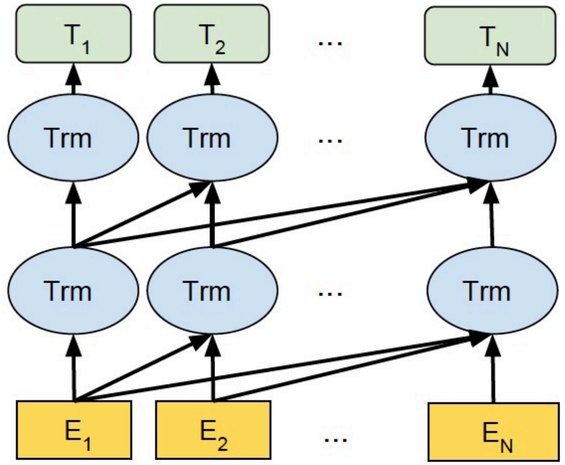
Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Generative Pre-Training

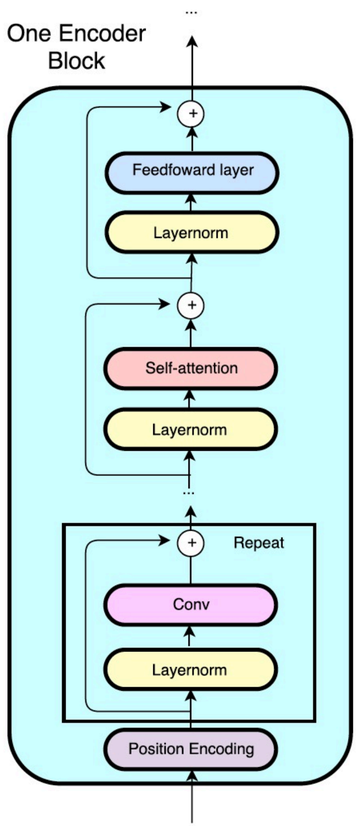
- 利用Transformer网络代替LSTM作为语言模型
  - 更好的捕获长距离语言结构
- 具体任务时使用有监督数据微调语言模型作为附属任务训练目标

Radford, A. & Salimans, T. Improving Language Understanding by Generative Pre-Training. (2018).

Generative Pre-Training



OpenAI GPT



Unsupervised pre-training

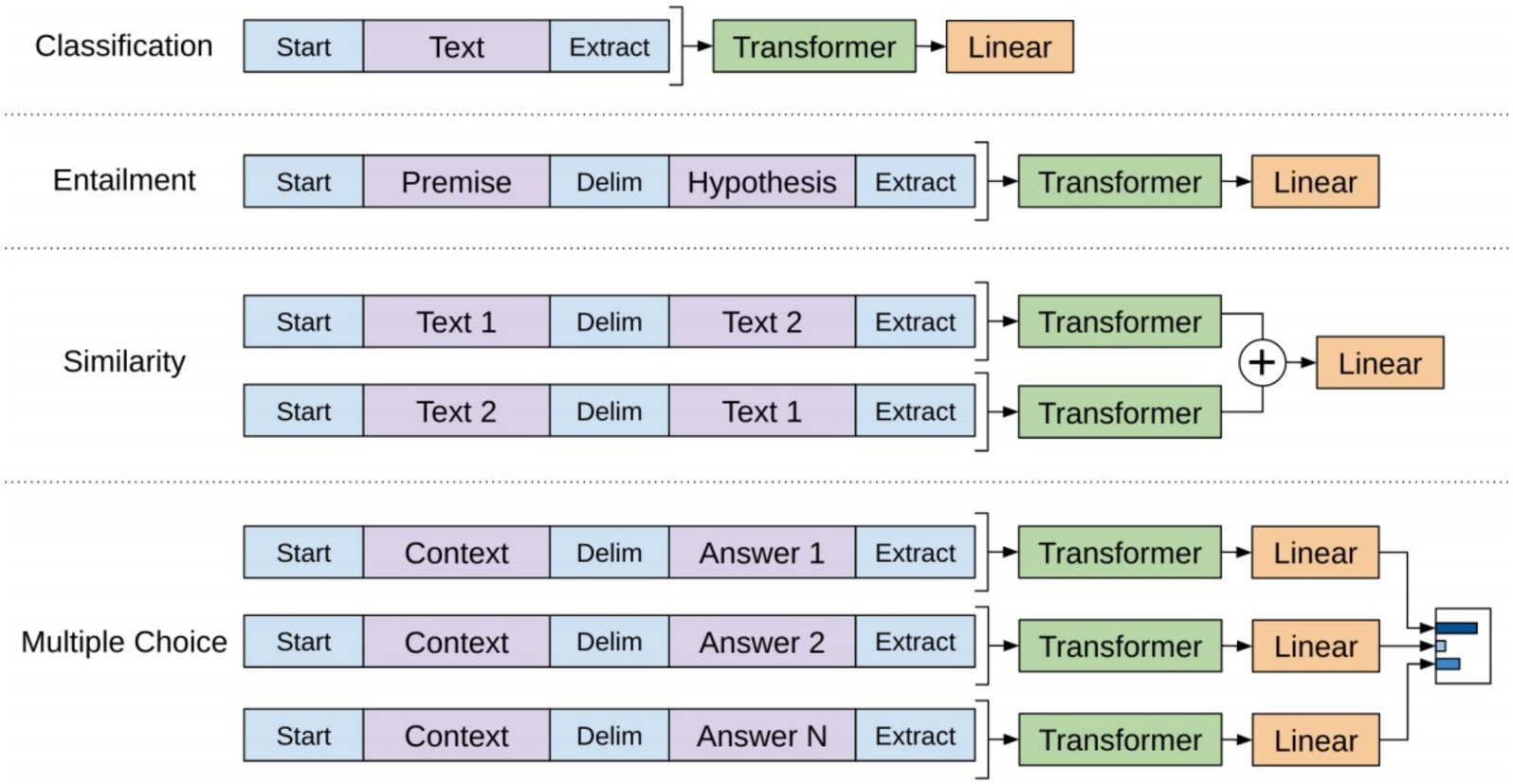
$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$
$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

Supervised fine-tuning

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

Radford, A. & Salimans, T. Improving Language Understanding by Generative Pre-Training. (2018).

Generative Pre-Training



Radford, A. & Salimans, T. Improving Language Understanding by Generative Pre-Training. (2018).

Generative Pre-Training

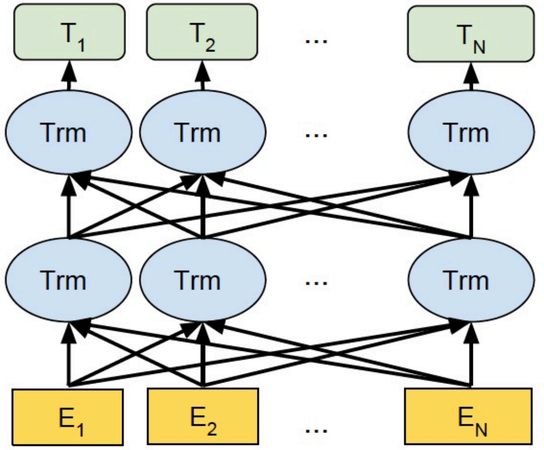
Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Radford, A. & Salimans, T. Improving Language Understanding by Generative Pre-Training. (2018).

Bidirectional Encoder Representations from Transformers

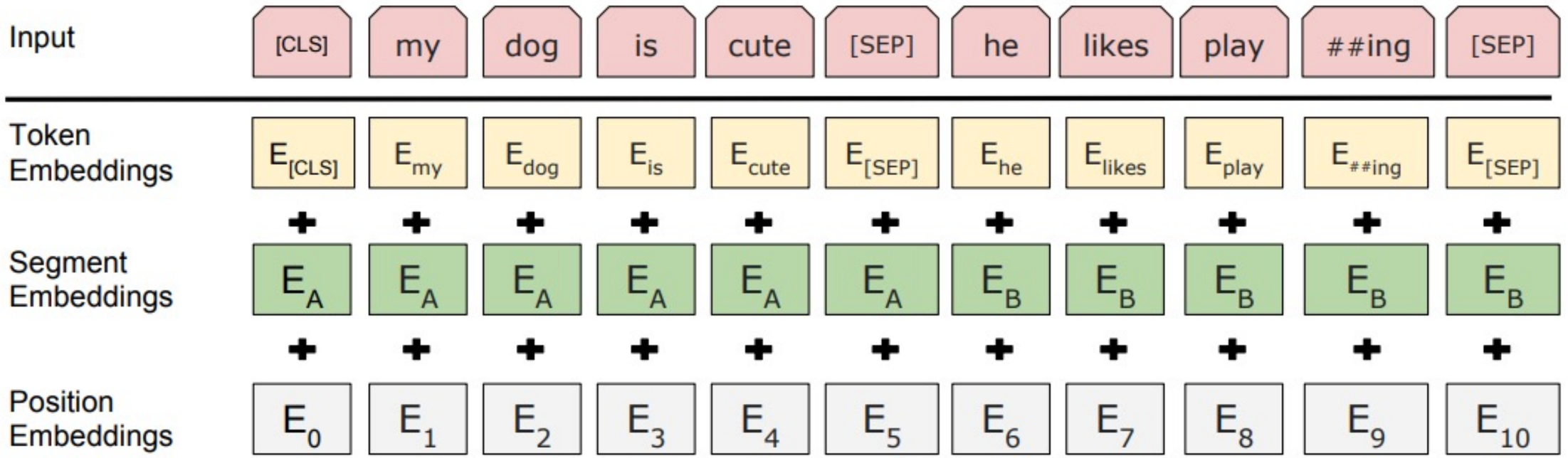
- 使用 Transformer 的编码器来作为语言模型，所有层中都是双向的
- 语言模型预训练
  - 遮挡语言模型 MLM
  - 预测下一个句子
- WordPiece Embedding 作为词向量
- 加入了位置向量和句子切分向量



BERT

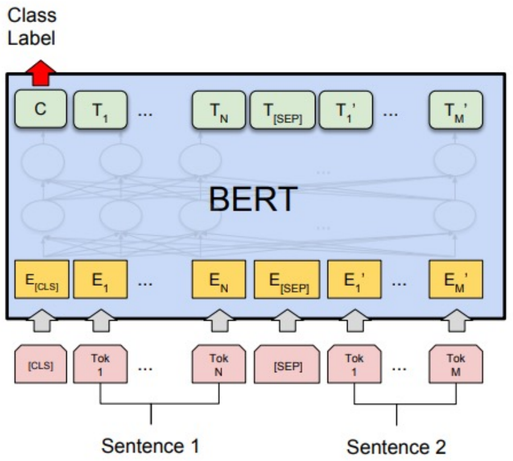
Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

Bidirectional Encoder Representations from Transformers

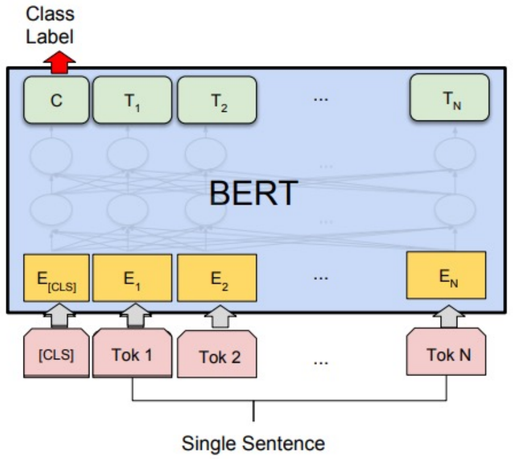


Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

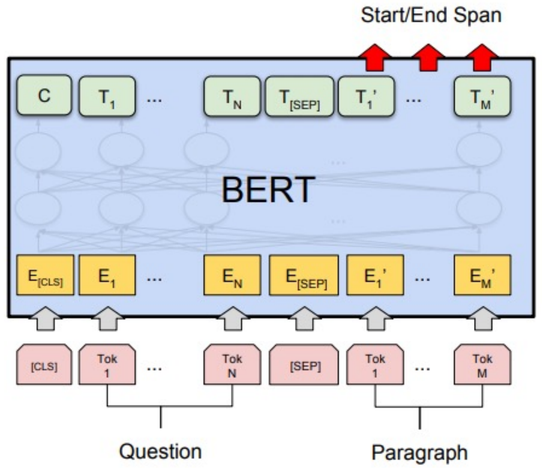
Bidirectional Encoder Representations from Transformers



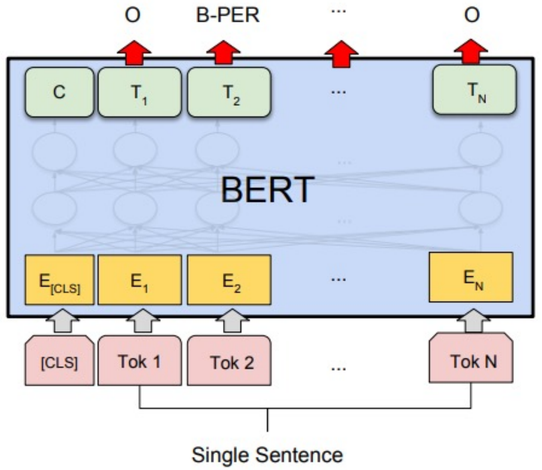
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE. SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

Bidirectional Encoder Representations from Transformers

Layers	Dev F1
Finetune All	96.4
First Layer (Embeddings)	91.0
Second-to-Last Hidden	95.6
Last Hidden	94.9
Sum Last Four Hidden	95.9
Concat Last Four Hidden	96.1
Sum All 12 Layers	95.5

Table 7: Ablation using BERT with a feature-based approach on CoNLL-2003 NER. The activations from the specified layers are combined and fed into a two-layer BiLSTM, without backpropagation to BERT.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 08, 2018	BERT (single model) *Google AI Language*	**80.005**	**83.061**
2 Nov 06, 2018	SLQA+BERT (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	77.003	80.209
3 Nov 08, 2018	BERT_base_aug (ensemble) *GammaLab*	76.721	79.611
4 Nov 05, 2018	MIR-MRC(F-Net) (single model) *Kangwon National University, Natural Language Processing Lab. & ForceWin, KP Lab.*	74.803	77.988

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

语言模型预训练

- 可以充分利用大规模的单语语料
- 一定程度上可以对一词多义进行建模
- 采用Pretrain 模式还可以在很大程度上缓解具体任务对模型结构的依赖

谢谢！