AI ProCon 开发者大会

CSDN

基于深度学习的机器阅读理解

Deep Learning Based Machine Reading Comprehension

崔一鸣

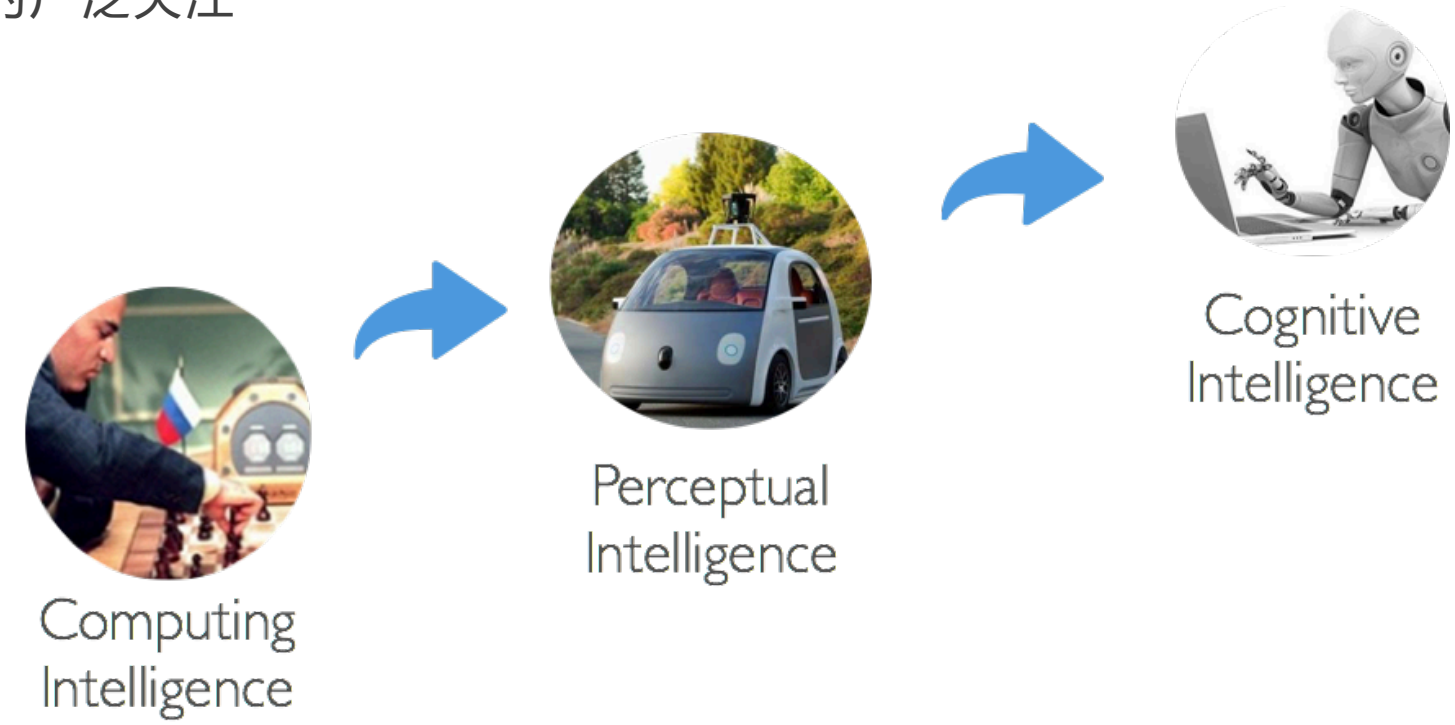
哈工大讯飞联合实验室(HFL)，科大讯飞

2019/9/7

机器阅读理解技术简介

机器阅读理解技术简介

- 人工智能的一个重要目标是**让机器能听会说，能理解会思考**
- 人工智能正处在从感知智能到认知智能跨越的时代
- **机器阅读理解（Machine Reading Comprehension, MRC）**作为认知智能的典型任务受到国内外研究人员的广泛关注

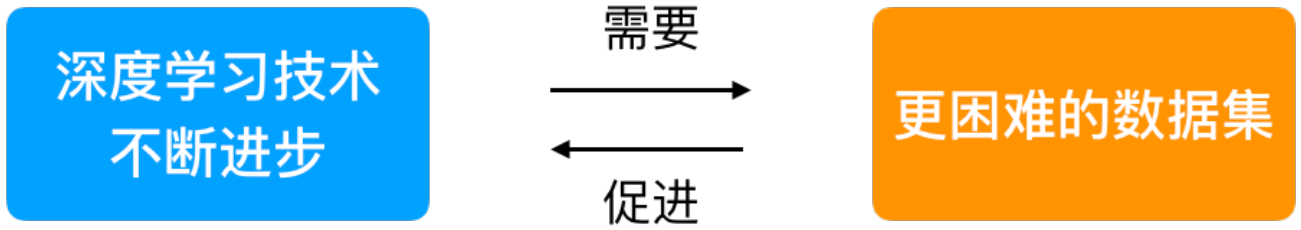


机器阅读理解技术简介

- **广义的阅读理解**
  - 指人类在认知世界的过程中逐步得以强化并对知识进行归纳、总结的能力
- **狭义的阅读理解**
  - 特指通过阅读文本文章，并且对相关问题进行解答的过程

机器阅读理解技术简介

- **为什么机器阅读理解在近些年来特别热门？**
  - NLP研究越来越依赖深度学习技术
  - 大规模机器机器阅读理解数据集的诞生
- **机器阅读理解的繁荣发展恰逢同时满足上述两个条件**



机器阅读理解技术简介：四要素

- **Document**
  - 需要机器阅读的篇章。根据篇章数量，分为单文档阅读理解、多文档阅读理解等
- **Question**
  - 根据篇章内容所提出的问题。根据问题类型，分为填空型、用户提问型等
- **Candidate**
  - 候选答案。根据任务类型，有时会与偶一些候选答案，如选择型阅读理解等
- **Answer**
  - 最终答案。根据任务类型，可能是单个词、篇章片段、生成的句子等

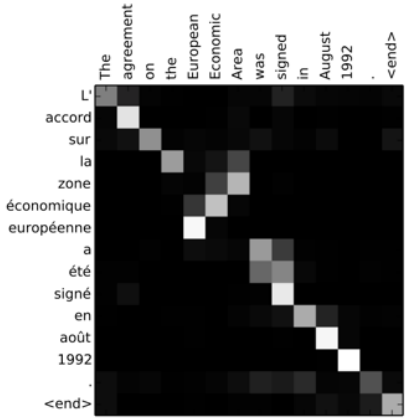
RACE
Passage Is it important to have breakfast every day? A short time ago, a test was given in the United States. People of different ages, from 12 to 83, were asked to have a test. During the test, these people were given all kinds of breakfast, and sometimes they got no breakfast at all. ...
Question What do the results show?
Candidates **A** ***They show that breakfast has affected on work and study.*** **B** Breakfast has little to do with a person's work. **C** A person will work better if he only has fruit and milk. **D** They show that girl students should have less for breakfast.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. ***EMNLP 2017****.*

机器阅读理解技术简介

- **主要探寻如何更好的建模“篇章”与“问题”之间的关系**
  - 一个很自然的方法是利用注意力机制（Attention）
  - 注意力机制的诞生源于计算机视觉领域
  - 2014年，Bengio等人首次应用注意力机制解决机器翻译问题

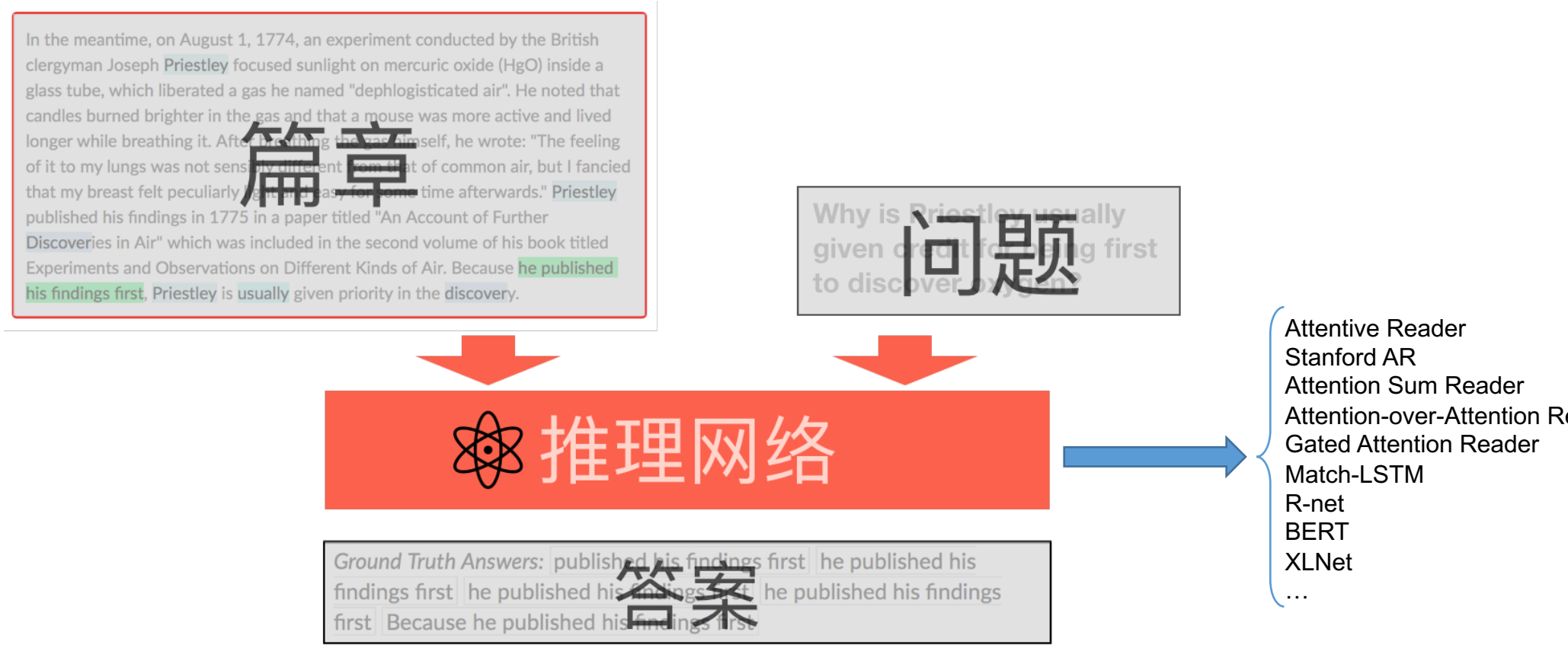
▲CV领域中的注意力机制



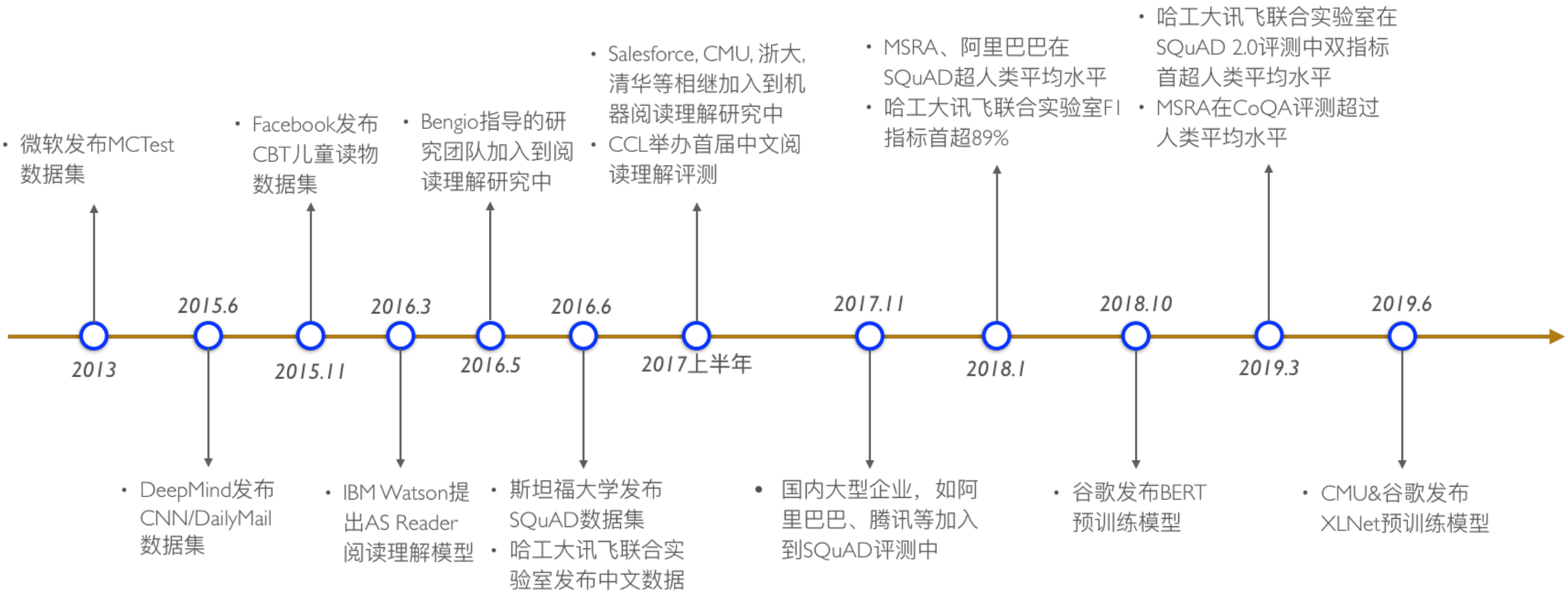
▲ NLP领域中的注意力机制

机器阅读理解技术简介

- 机器阅读理解相关神经网络模型层出不穷



机器阅读理解技术简介


• 微软发布MCTest数据集
2013
2015.6
• DeepMind发布CNN/DailyMail数据集
• Facebook发布CBT儿童读物数据集
2015.11
2016.3
• IBM Watson提出AS Reader阅读理解模型
• Bengio指导的研究团队加入到阅读理解研究中
2016.5
2016.6
• 斯坦福大学发布SQuAD数据集
• 哈工大讯飞联合实验室发布中文数据
• Salesforce, CMU, 浙大, 清华等相继加入到机器阅读理解研究中
• CCL举办首届中文阅读理解评测
2017上半年
2017.11
• 国内大型企业，如阿里巴巴、腾讯等加入到SQuAD评测中
• MSRA、阿里巴巴在SQuAD超人类平均水平
• 哈工大讯飞联合实验室F1指标首超89%
2018.1
2018.10
• 谷歌发布BERT预训练模型
• 哈工大讯飞联合实验室在SQuAD 2.0评测中双指标首超人类平均水平
• MSRA在CoQA评测超过人类平均水平
2019.3
2019.6
• CMU&谷歌发布XLNet预训练模型

传统机器阅读理解技术

填空型阅读理解

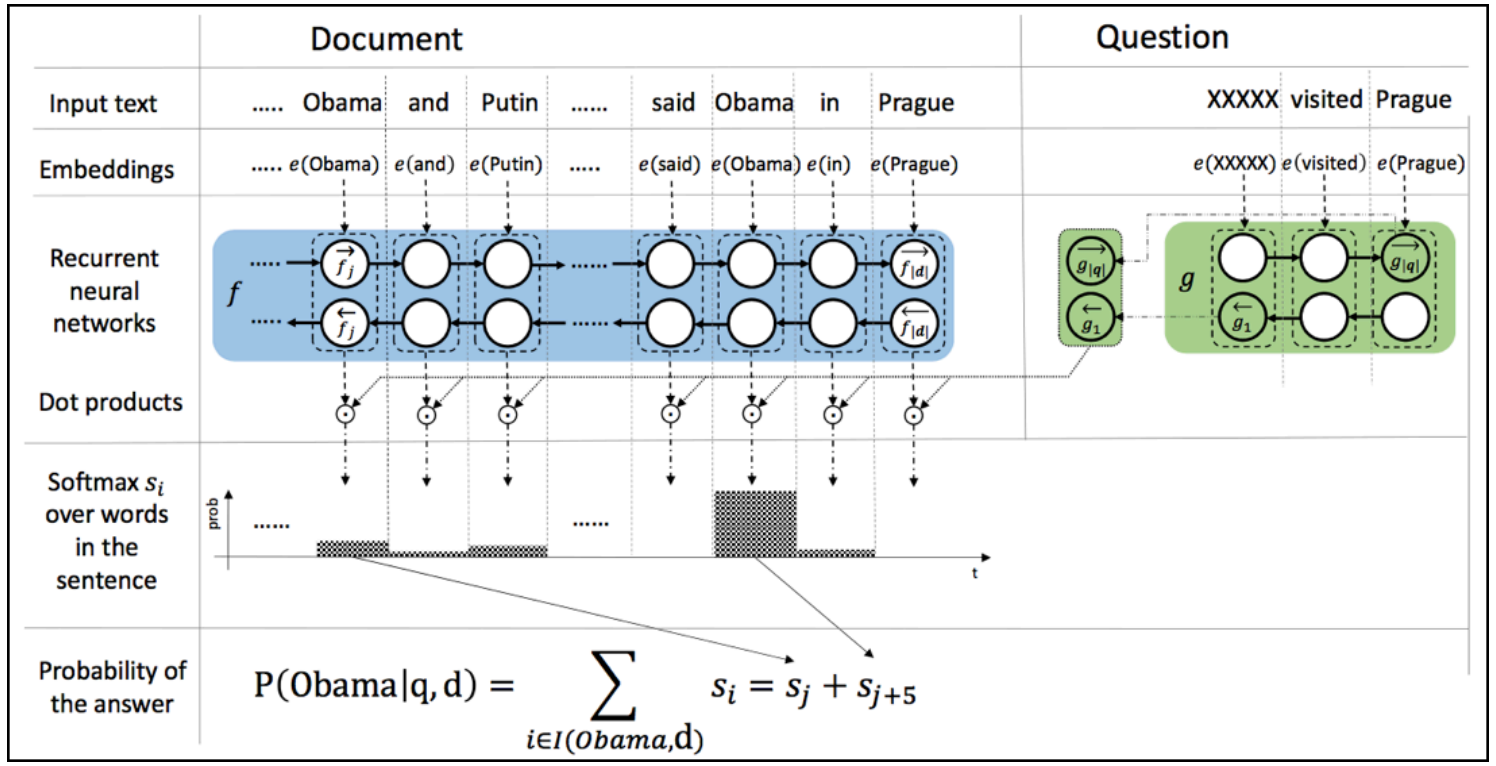
- **填空型阅读理解（Cloze-style RC）任务是近年机器阅读理解流行的开端**
  - 根据上下文的识别和判断，对空缺中的词进行填补
  - 主要考察对名词、名实体类别的填空
  - 限制使用文中出现的单词进行填空

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom. Teaching Machines to Read and Comprehend. ***NIPS 2015.***

AS Reader

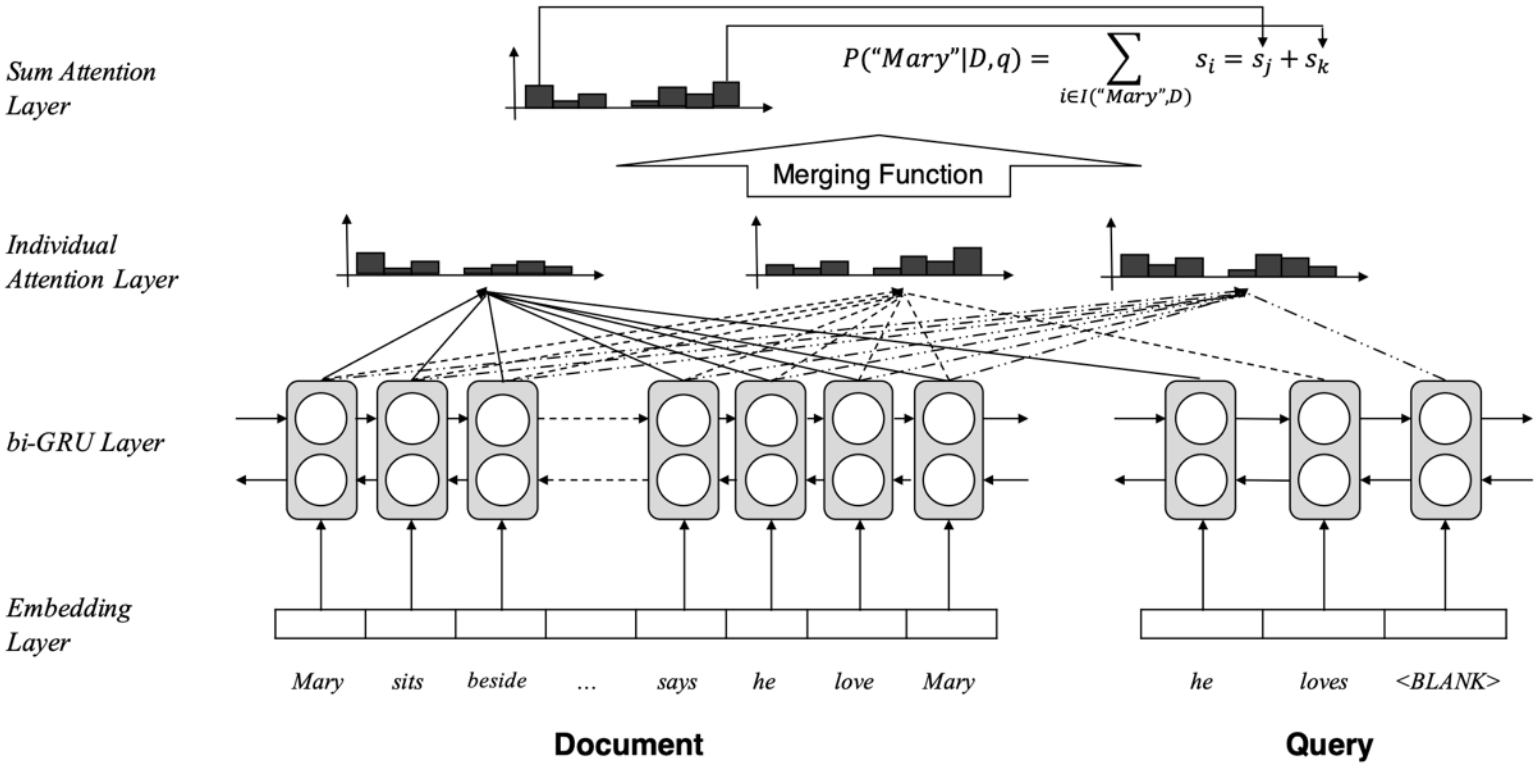
- **Attention Sum Reader (ACL 2016, IBM Research)**
  - 利用Pointer Network直接对空缺词进行预测，极大的简化了该任务的处理流程



Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, Jan Kleindienst. Text Understanding with the Attention Sum Reader Network. ***ACL 2016****.*

CAS Reader

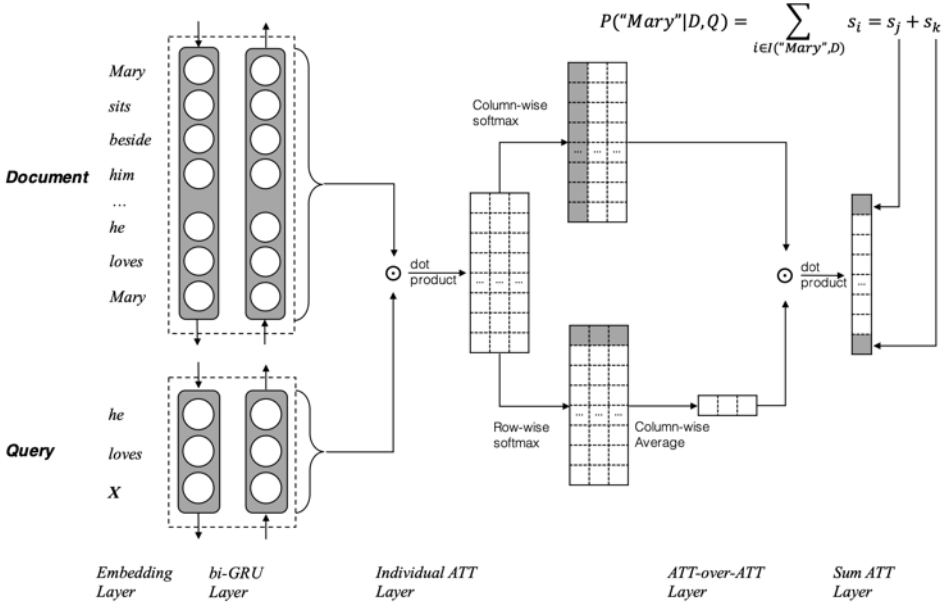
- **Consensus Attention-based Sum Reader (COLING 2016)**
  - 在AS Reader的基础上进一步细化了对问题的建模



Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, Guoping Hu. Consensus Attention-based Neural Networks for Chinese Reading Comprehension. ***COLING 2016****.*

AoA Reader

- **Attention-over-Attention Reader (ACL 2017)**
  - 提出层叠式注意力机制，双向建模篇章与问题之间的关系
  - AoA Reader在多个阅读理解任务中获得显著性能提升
  - 使用灵活，可作为常规的Attention计算，在后续工作中得到广泛应用



	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann et al., 2015)	55.0	57.0	-	-	-	-
Attentive Reader (Hermann et al., 2015)	61.6	63.0	-	-	-	-
Human (context+query) (Hill et al., 2015)	-	-	-	*81.6*	-	*81.6*
MemNN (window + self-sup.) (Hill et al., 2015)	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec et al., 2016)	68.6	69.5	73.8	68.6	68.8	63.4
CAS Reader (Cui et al., 2016)	68.2	70.0	74.2	69.2	68.2	65.7
Stanford AR (Chen et al., 2016)	72.4	72.4	-	-	-	-
GA Reader (Dhingra et al., 2016)	73.0	73.8	74.9	69.0	69.0	63.9
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	75.2	68.6	72.1	69.2
EpiReader (Trischler et al., 2016)	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader	73.1	**74.4**	**77.8**	**72.0**	**72.2**	**69.4**
AoA Reader + Reranking	-	-	**79.6**	**74.0**	**75.7**	**73.1**
MemNN (Ensemble)	66.2	69.4	-	-	-	-
AS Reader (Ensemble)	73.9	75.4	74.5	70.6	71.1	68.9
GA Reader (Ensemble)	76.4	77.4	75.5	71.9	72.1	69.4
EpiReader (Ensemble)	-	-	76.6	71.8	73.6	70.6
Iterative Attention (Ensemble)	74.5	75.7	76.9	72.0	74.1	**71.0**
AoA Reader (Ensemble)	-	-	**78.9**	**74.5**	**74.7**	70.8
AoA Reader (Ensemble + Reranking)	-	-	**80.3**	**75.6**	**77.0**	**74.1**

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, Guoping Hu. Attention-over-Attention Neural Networks for Reading Comprehension. ***ACL 2017****.*

阅读理解技术的交叉应用

- **利用阅读理解模型解决中文零指代问题**
  - 提出制造“伪数据”的方法来解决领域内数据少的问题
  - 应用“**大数据+小数据**”的模式来解决领域迁移问题
  - 首次将阅读理解模型应用于其他NLP任务并取得显著提升

[原句]
鸭子看见了，以为是一条鱼，赶紧游过去捉。

[零指代消解后]
鸭子看见了，（鸭子）以为是一条鱼，鸭子赶紧游过去捉

▲ 中文零指代任务示例

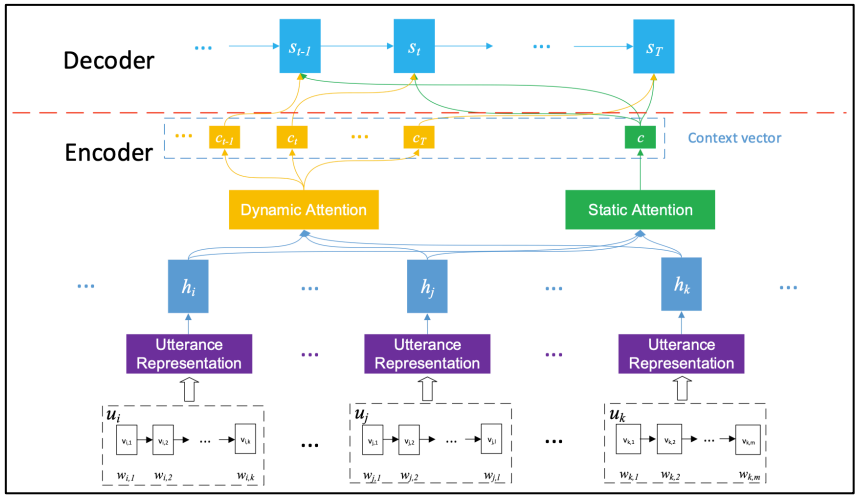
	NW (84)	MZ (162)	WB (284)	BN (390)	BC (510)	TC (283)	**Overall**
Kong and Zhou (2010)	34.5	32.7	45.4	51.0	43.5	48.4	44.9
Chen and Ng (2014)	38.1	31.0	50.4	45.9	53.8	**54.9**	48.7
Chen and Ng (2015)	46.4	39.0	51.8	53.8	49.4	52.7	50.2
Chen and Ng (2016)	48.8	41.5	56.3	**55.4**	50.8	53.1	52.2
Our Approach†	**59.2**	**51.3**	**60.5**	53.9	**55.5**	52.9	**55.3**

▲ OntoNotes 5.0上效果显著

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****.*

阅读理解技术的交叉应用

- **利用阅读理解模型解决开放域对话系统问题**
  - 阅读理解的推广即是上下文的理解
  - 我们首次将阅读理解的思想应用于开放域对话系统中
  - 提出基于静态和动态的注意力机制，有效建模历史对话流



▲模型结构

Models	Ubuntu			OpenSubtitles		
	Average	Greedy	Extrema	Average	Greedy	Extrema
LSTM	0.2300	0.1689	0.1574	0.5549	0.5029	0.3897
HRED	0.5770	0.4169	0.3914	0.5571	0.5033	0.3932
VHRED	0.5419	0.3839	0.3627	0.5248	0.4821	0.3556
CVAE	0.5672	0.3982	0.3689	0.4708	0.3390	0.3173
WSI	0.5775	0.4196	0.3893	0.5598	0.4964	0.3903
HRAN	0.5964	0.4139	0.3898	0.5617	0.5195	0.3898
Dynamic$_{\rightleftarrows}$	0.5750	0.4043	0.3802	0.5487	0.5054	0.3812
Dynamic$_{\rightarrow}$	0.5968	0.4132	0.3877	0.5629	0.5193	0.3905
Static$_{\rightleftarrows}$	0.5998	0.4124	0.3886	0.5475	0.5147	0.3862
Static$_{\rightarrow}$	**0.6121**†	**0.4293**†	**0.3975**†	**0.5656**†	**0.5232**†	**0.3937**†

▲英文数据上效果显著

Wei-Nan Zhang, Yiming Cui, Yifa Wang, Qingfu Zhu, Lingzhi Li, Lianqiang Zhou, Ting Liu..Context-Sensitive Generation of Open-Domain Conversational Responses. ***COLING 2018****.*

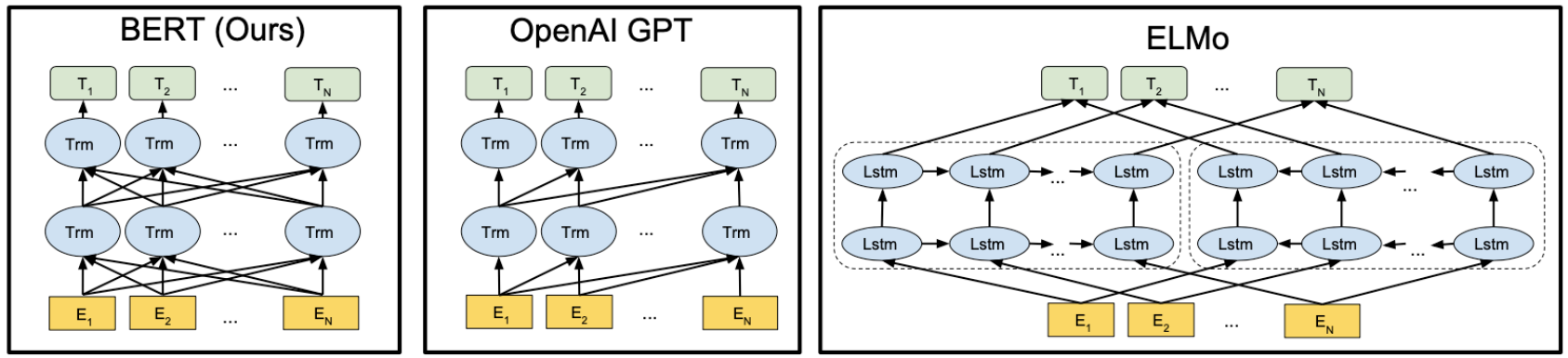
基于预训练模型的机器阅读理解技术

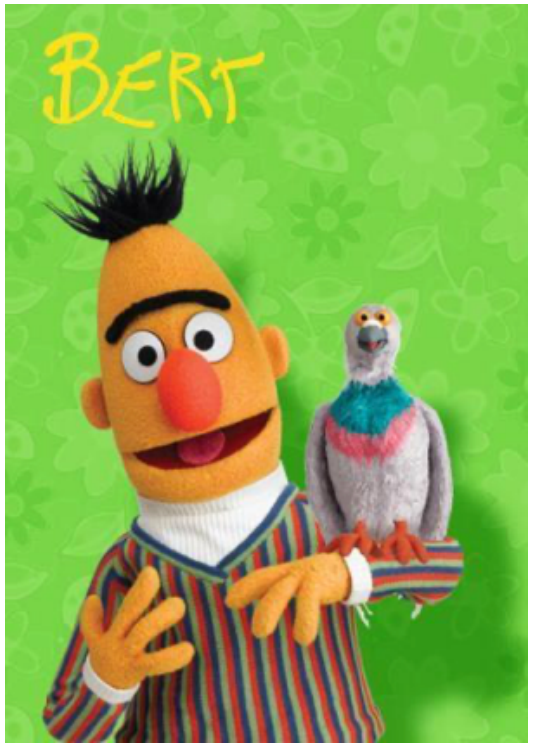
基于预训练模型的机器阅读理解技术

- **预训练模型通常是大规模数据上训练的模型，包含丰富的语义表示**
  - 在CV领域，以ImageNet为代表的预训练模型早已广泛流行
  - 在NLP领域，从去年下半年开始，预训练模型逐渐受到关注
  - AI2发布了ELMo模型
  - Google发布了BERT模型
  - 百度发布了ERNIE模型
  - CMU/Google发布了XLNet模型
  - Facebook发布了RoBERTa模型
  - ……

BERT

- **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL 2019, Google)**
  - NAACL 2019 最佳论文
  - 17,000+ Star; 4,500+ Fork @ GitHub
  - 开启了NLP任务的新范式：预训练+精调



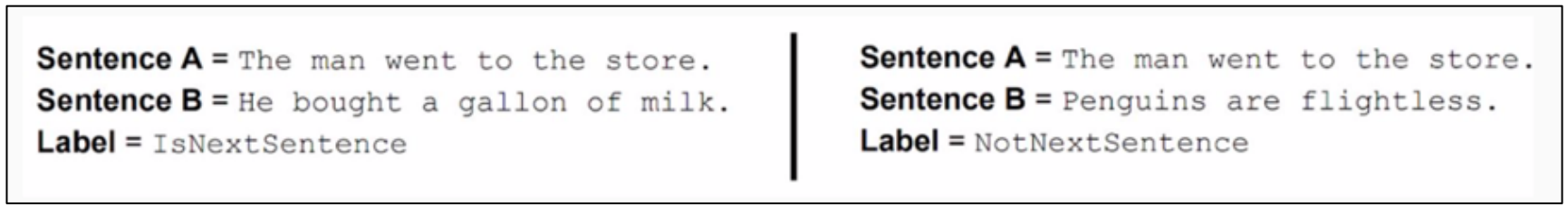


Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ***NAACL 2019****.*

BERT

- **任务一：Masked LM (MLM)**
  - 15%几率被Mask，其中：80%替换为[MASK]，10%替换为随机词，10%保留原词

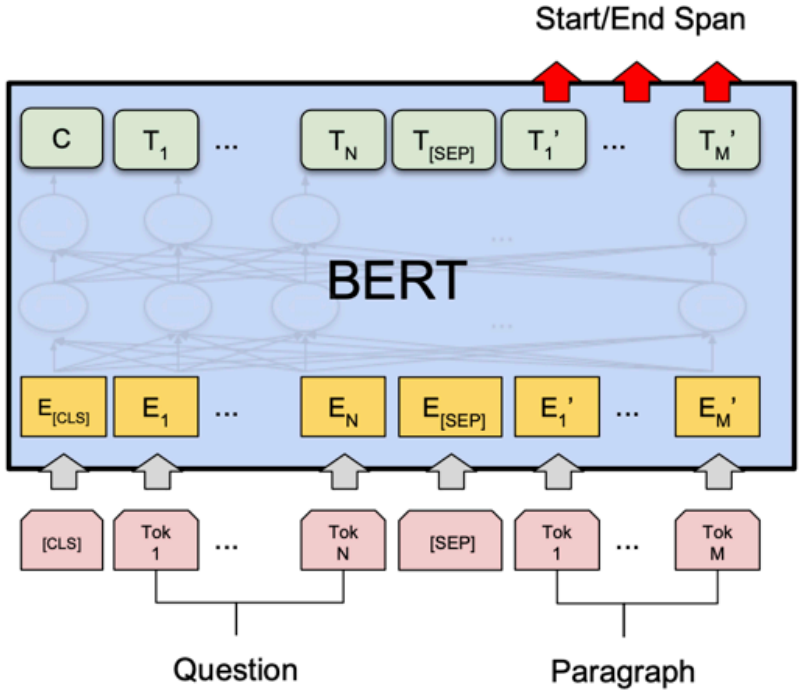
- **任务二：Next Sentence Prediction (NSP)**
  - 判断句子B是否是句子A后一个句子



Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ***NAACL 2019****.*

BERT

- **只需指定输入输出就能得到相应任务模型**
  - SQuAD任务：只需将问题和篇章拼接送入BERT，输出start/end的指针
  - BERT模型横扫以往的所有模型，在SQuAD任务中达到state-of-the-art效果

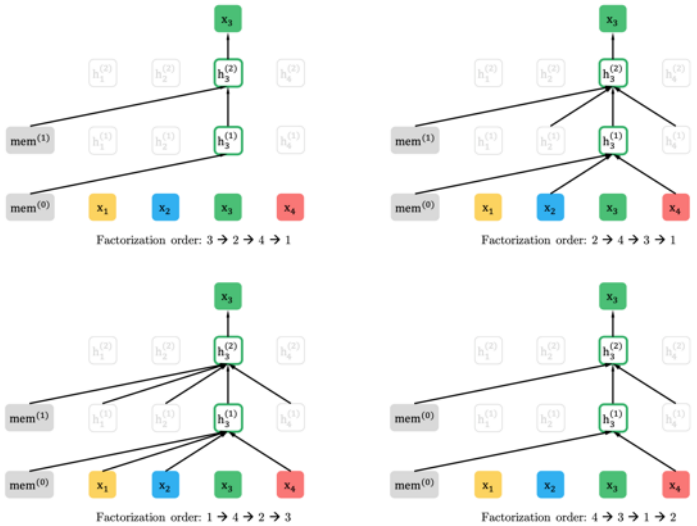


System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ***NAACL 2019****.*

XLNet

- **XLNet: Generalized Autoregressive Pretraining for Language Understanding**
  - 提出一种泛化自回归预训练方法，克服了BERT的“预训练-精调”之间的差异
  - 创新性的提出了双流自注意力结构（Two-Stream Self-Attention）
  - 进一步刷新SQuAD结果，大幅超过BERT相关模型

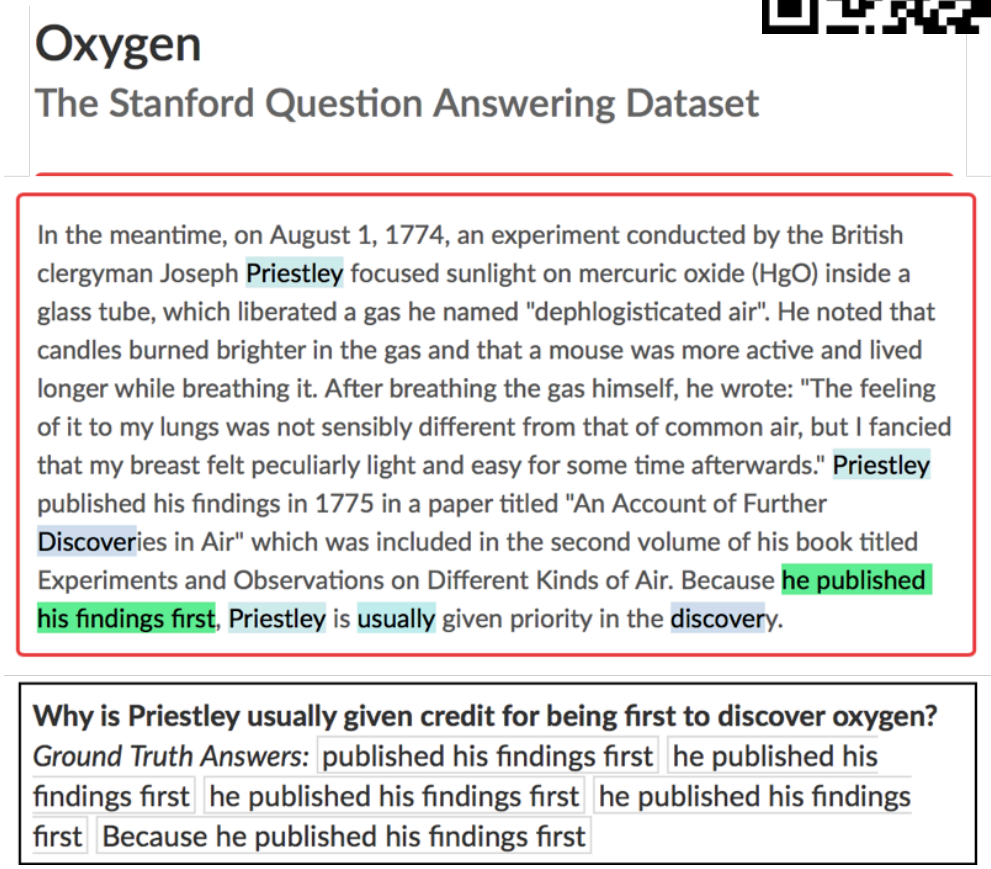


SQuAD1.1	EM	F1	SQuAD2.0	EM	F1
Dev set results without data augmentation					
BERT [10]	84.1	90.9	BERT† [10]	78.98	81.77
XLNet	**88.95**	**94.52**	XLNet	**86.12**	**88.79**
Test set results on leaderboard, with data augmentation (as of June 19, 2019)					
Human [27]	82.30	91.22	BERT+N-Gram+Self-Training [10]	85.15	87.72
ATB	86.94	92.64	SG-Net	85.23	87.93
BERT* [10]	87.43	93.16	BERT+DAE+AoA	85.88	88.62
XLNet	**89.90**	**95.08**	XLNet	**86.35**	**89.13**

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. ***arXiv pre-print.***

SQuAD

- **斯坦福大学发布Stanford Question Answering Dataset (SQuAD) 数据集**
  - 开启了阅读理解数据集的新篇章
  - 答案不再只是单个词，扩展到短语甚至是句子，难度更大
  - 问题不再是自动生成，采用了人工标注，问题更加自然
  - 相比之前的人工标注集合，规模较大（10万问题）
- **篇章**：维基百科文章，切分成若干段落组成“小篇章”
- **问题**：由人工提问，问题类型较多，丰富了问题的多样性(what/when/where/who/how/why等)
- **答案**：篇章中某个连续片段，搜索空间更大，难度更高



Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. ***EMNLP 2016****.*

BERT+DAE+AoA

- **BERT+DAE+AoA**
- **拆分技术一：BERT**
  - 采用最新的文本语义表示模型BERT，丰富文本的表示
- **拆分技术二： DAE**
  - 借鉴早期在中文零指代任务中的技术模式，采用自动的方法生成大量的训练“伪数据”，进一步挖掘<篇章，问题，答案>之间的关系
- **拆分技术三： AoA**
  - 团队持续积累的AoA系列技术得到进一步提升
  - 对历史注意力进行迭代建模，进一步提高注意力机制的准确性

SQuAD 2.0夺冠

- 2019年3月，哈工大讯飞联合实验室提交的模型在业界首次双指标全部超过人类平均水平

Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**82.374**	**85.310**
2 Nov 16, 2018	AoA + DA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	81.178	84.251
3 Nov 16, 2018	Candi-Net+BERT (single model) *42Maru NLP Team*	80.106	82.862
3 Nov 08, 2018	BERT (single model) *Google AI Language*	80.005	83.061
4 Nov 09, 2018	L6Net + BERT (single model) *Layer 6 AI*	79.181	82.259
5 Nov 06, 2018	SLQA+BERT (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	77.003	80.209

▲2018年11月榜单

Leaderboard

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

▲2019年3月20日榜单

SQuAD2.0
The Stanford Question Answering Dataset

Pranav Rajpurkar @pranavrajpur... · 5天

Human Performance on SQuAD2.0 has now been surpassed! Been in the making for the last year, but finally achieved! Congratulations Joint Laboratory of HIT and iFLYTEK Research on this milestone! @KCrosner @stanfordnlp

▲SQuAD发起人表示祝贺

对话型机器阅读理解

- **对话型机器阅读理解：阅读理解 + 多轮对话**
  - 阅读理解任务与其他NLP任务的交叉研究
  - 相比一问一答的形式，通过多轮对话完成阅读理解过程是自然的
  - 2018年8月，AI2&华盛顿大学&斯坦福大学等单位发布QuAC数据集
  - 同月，斯坦福大学发布了CoQA数据集
- **任务难点**
  - 标准答案依赖于但不再是篇章中的某个连续片段
  - 多轮对话中的指代消解对问题理解提出新的挑战

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Siva Reddy, Danqi Chen, Christopher D. Manning. CoQA: A Conversational Question Answering Challenge. ***TACL 2019.***

对话型机器阅读理解

- 哈工大讯飞联合实验室在对话型阅读理解评测CoQA、QuAC中夺冠

Leaderboard

Rank	Model	In-domain	Out-of-domain	Overall
	Human Performance *Stanford University* (Reddy et al. '18)	89.4	87.4	88.8
1 Dec 12, 2018	D-AoA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	**81.4**	**77.3**	**80.2**
2 Nov 29, 2018	SDNet (ensemble model) *Microsoft Speech and Dialogue Research Group* https://arxiv.org/abs/1812.03593	80.7	75.9	79.3
3 Nov 26, 2018	SDNet (single model) *Microsoft Speech and Dialogue Research Group* https://arxiv.org/abs/1812.03593	78.0	73.1	76.6
4 Oct 06, 2018	FlowQA (single model) *Allen Institute for Artificial Intelligence* https://arxiv.org/abs/1810.06683	76.3	71.8	75.0
5 Dec 10, 2018	BiDAF++ (single model) *Beijing University of Posts and Telecommunications*	71.1	65.5	69.5
6 Sep 27, 2018	BiDAF++ (single model) *Allen Institute for Artificial Intelligence* https://arxiv.org/abs/1809.10735	69.4	63.8	67.8

▲ 2018年12月，荣登CoQA评测榜首

Leaderboard

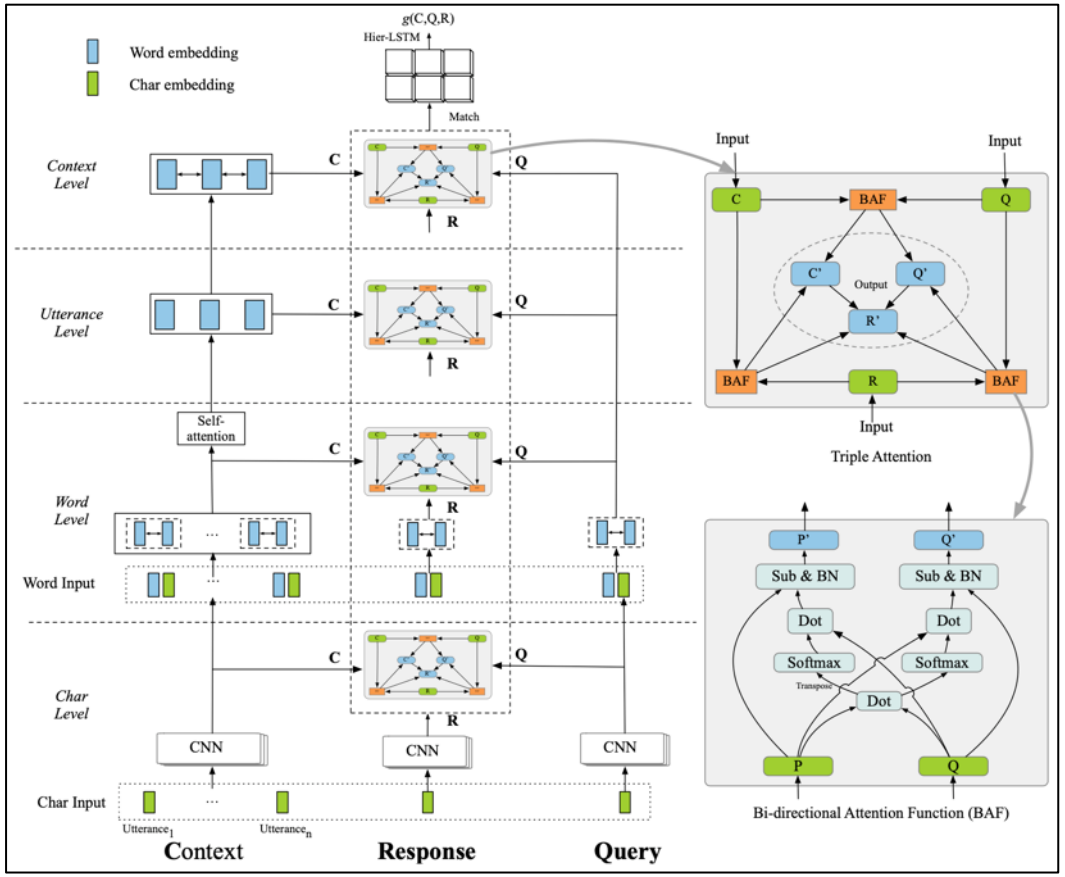
There can be only one duck.

Rank	Model	F1	HEQQ	HEQD
	Human Performance (Choi et al. EMNLP '18)	81.1	100	100
Mar 14, 2019	ConvBERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	**68.0**	**63.5**	**9.1**
2 Mar 14, 2019	Bert-FlowDelta (single model) *Anonymous*	66.1	61.0	7.4
3 Mar 7, 2019	BERT w/ 2-context (single model) *NTT Media Intelligence Labs*	64.9	60.2	6.1
4 Feb 21, 2019	GraphFlow (single model) *Anonymous*	64.9	60.3	5.1
5 Sep 26, 2018	FlowQA (single model) *Allen Institute of AI* https://arxiv.org/abs/1810.06683	64.1	59.6	5.8

▲ 2019年3月，荣登QuAC评测榜首

对话型机器阅读理解

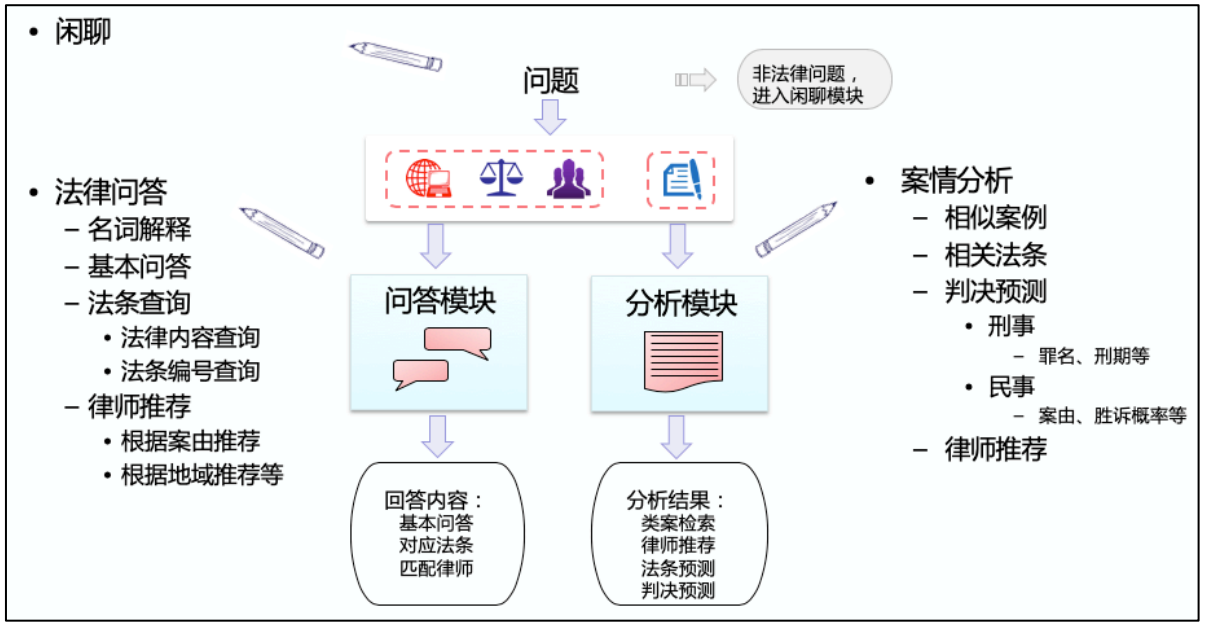
- **夺冠系统重点技术剖析**
  - 对话要素间的建模
    - 充分建模历史对话流、当前Query、回复之间的关系
  - 层次化的对话流建模
    - 从word到utterance再到context层级，进行层次化建模，充分考虑各粒度之间的联系
  - 利用最新的文本语义表示模型BERT，丰富文本的语义表示，提升整体效果
- **同时，分拆技术应用于多轮对话任务，在中文Douban、英文Ubuntu数据上获得显著性能提升**



Wentao Ma, Yiming Cui, Nanshao Su He, Wei-Nan Zhang, Ting Liu, Shijin Wang, Guoping Hu. TripleNet: Triple Attention Network for Multi-Turn Response Selection in Retrieval-based Chatbots. ***CoNLL 2019.***

对话型机器阅读理解：应用——法小飞

- **哈工大讯飞联合实验室推出“法小飞”法律咨询助手**
  - 2018年5月，哈工大讯飞联合实验室正式向公众提供测试
  - “法小飞”基本功能：法律问答、案情分析、闲聊等
  - 在法律问答中，利用**对话型机器阅读理解技术**在人机交互的过程中提供更加精准的答案



Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, Ting Liu. IFlyLegal: A Chinese Legal System for Consultation, Law Searching, and Document Analysis. ***EMNLP 2019 Demo****.*

中文阅读理解数据

- **哈工大讯飞联合实验室、百度等单位相继发布中文阅读理解数据，促进该领域发展**
- **填空型阅读理解**
  - PD&CFT（COLING 2016，HFL）、CMRC 2017（LREC 2018，HFL）、WebQA（百度）
- **篇章片段抽取型阅读理解**
  - CMRC 2018（EMNLP 2019，HFL）、DRCD（台达研究院）
- **开放域阅读理解**
  - DuReader（ACL 2018 MRQA Workshop，百度）
- **其他阅读理解数据**
  - C^3（腾讯）、CMRC 2019（HFL）、ChID（ACL 2019，清华大学）、CJRC（CCL 2019，HFL）

中文预训练模型

- **哈工大讯飞联合实验室发布基于Whole Word Masking（WWM）的中文BERT**
  - 将谷歌提出的Whole Word Masking技术应用在中文中，使用大规模语料进行预训练
  - 在多个NLP数据集上获得显著性能提升，相关资源已开源供学术研究及参考

[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]** 。

▲ Whole Word Masking 技术

模型	开发集	测试集	挑战集
BERT	65.5 (64.4) / 84.5 (84.0)	70.0 (68.7) / 87.0 (86.3)	18.6 (17.0) / 43.3 (41.3)
ERNIE	65.4 (64.3) / 84.7 (84.2)	69.4 (68.2) / 86.6 (86.1)	19.6 (17.0) / 44.3 (42.8)
BERT-wwm	66.3 (65.0) / 85.6 (84.7)	70.5 (69.1) / 87.4 (86.7)	21.0 (19.3) / 47.0 (43.9)
BERT-wwm-ext	**67.1 (65.6) / 85.7 (85.0)**	**71.4 (70.0) / 87.7 (87.0)**	**24.0 (20.0) / 47.3 (44.6)**

▲ CMRC 2018 简体中文阅读理解

模型	开发集	测试集
BERT	83.1 (82.7) / 89.9 (89.6)	82.2 (81.6) / 89.2 (88.8)
ERNIE	73.2 (73.0) / 83.9 (83.8)	71.9 (71.4) / 82.5 (82.3)
BERT-wwm	84.3 (83.4) / 90.5 (90.2)	82.8 (81.8) / 89.7 (89.0)
BERT-wwm-ext	**85.0 (84.5) / 91.2 (90.9)**	**83.6 (83.0) / 90.4 (89.9)**

▲ DRCD 繁体中文阅读理解

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. Pre-Training with Whole Word Masking for Chinese BERT. ***arXiv pre-print****.*

中文预训练模型

- **哈工大讯飞联合实验室发布中文XLNet预训练模型**
  - 探究XLNet在中文下的表现，填补该模型在中文领域的空白
  - 在中文（简体、繁体）阅读理解任务上获得显著性能提升，超过BERT相关模型
  - 相关预训练模型已开源供学术研究及参考

模型	开发集	测试集	挑战集
BERT	65.5 (64.4) / 84.5 (84.0)	70.0 (68.7) / 87.0 (86.3)	18.6 (17.0) / 43.3 (41.3)
BERT-wwm	66.3 (65.0) / 85.6 (84.7)	70.5 (69.1) / 87.4 (86.7)	21.0 (19.3) / 47.0 (43.9)
BERT-wwm-ext	**67.1** (65.6) / 85.7 (85.0)	**71.4 (70.0)** / 87.7 (87.0)	24.0 (20.0) / 47.3 (44.6)
XLNet-mid	66.8 **(66.3) / 88.4 (88.1)**	69.3 (68.5) / **89.2 (88.8)**	**29.1 (27.1) / 55.8 (54.9)**

▲ CMRC 2018 简体中文阅读理解

模型	开发集	测试集
BERT	83.1 (82.7) / 89.9 (89.6)	82.2 (81.6) / 89.2 (88.8)
BERT-wwm	84.3 (83.4) / 90.5 (90.2)	82.8 (81.8) / 89.7 (89.0)
BERT-wwm-ext	85.0 (84.5) / 91.2 (90.9)	83.6 (83.0) / 90.4 (89.9)
XLNet-mid	**85.3 (84.9) / 93.5 (93.3)**	**85.5 (84.8) / 93.6 (93.2)**

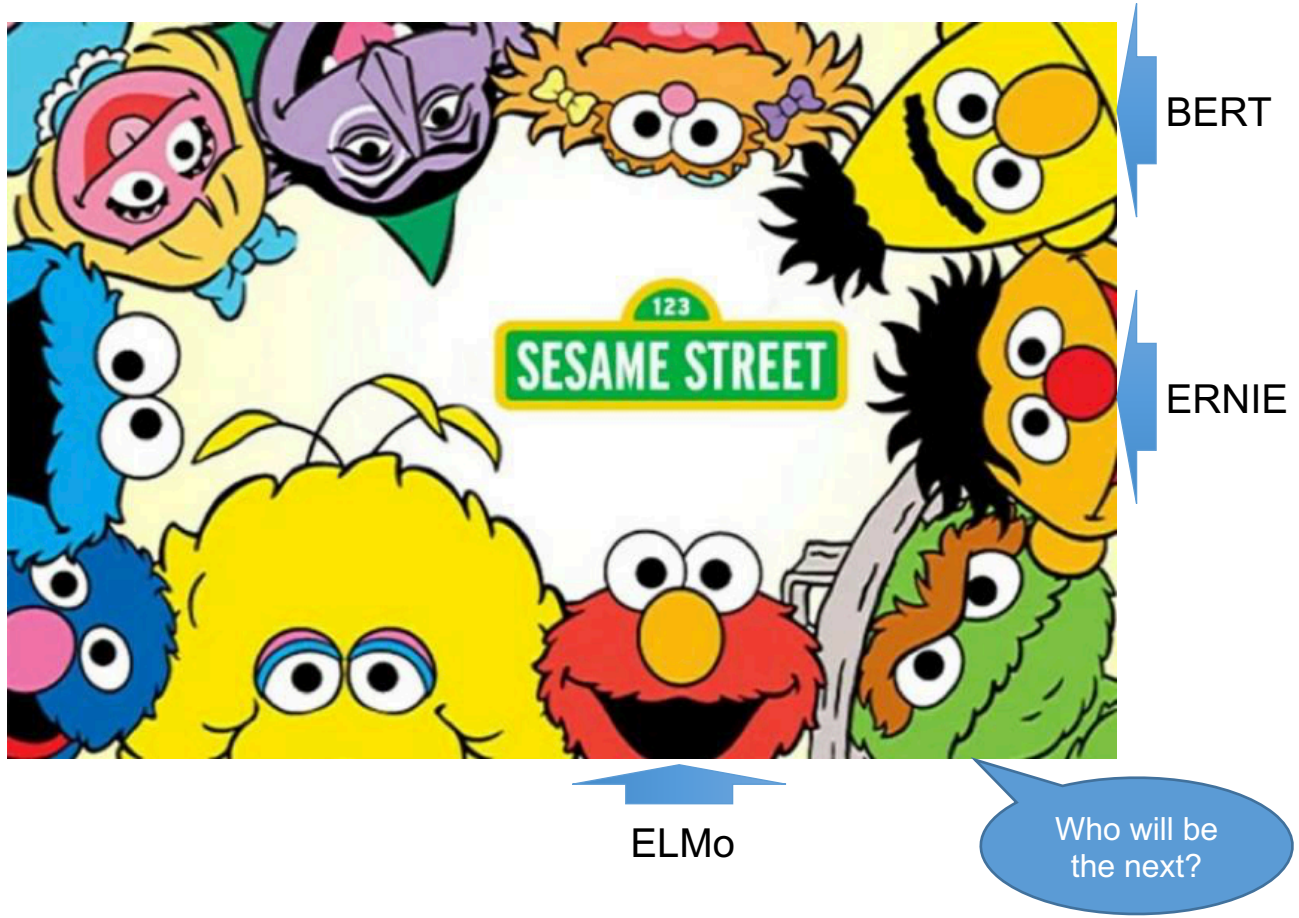
▲ DRCD 繁体中文阅读理解

Yiming Cui. https://github.com/ymcui/Chinese-PreTrained-XLNet

技术研究热点及趋势

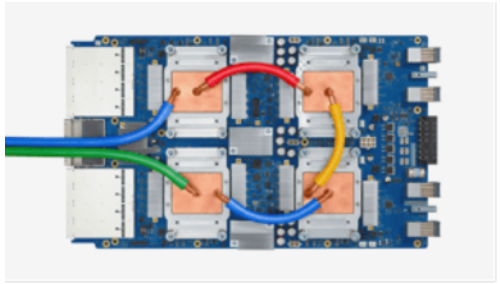
技术发展热点（1）：预训练

- **预训练模型已成为NLP领域新基底**
  - BERT (Google)
  - XLNet (CMU / Google)
  - SpanBERT (Facebook)
  - RoBERTa (Facebook)
  - ...
- **中文预训练模型**
  - ERNIE (Baidu)
  - Chinese BERT-wwm (HFL)
  - Chinese XLNet (HFL)



技术发展热点（1）：预训练

- **技术发展现状及问题**
  - “马太效应”愈加凸显，是不是又回到“拼算力”的时代？
  - 模型参数越来越多，训练和预测速度大大降低
- **未来的研究趋势**
  - 探索效果、效率更优的预训练技术
  - Task-Specific预训练模型技术的探索
  - 预训练模型的压缩和裁剪，提高训练和预测效率



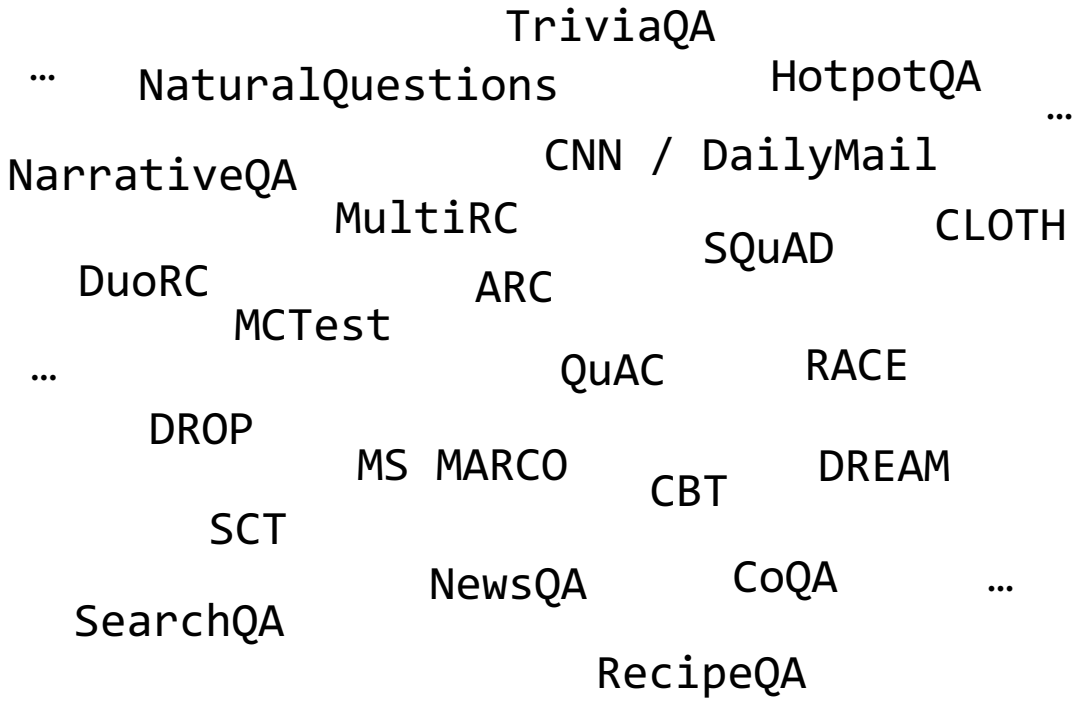
▲Google Cloud TPU v3
420TFLOPS / 128GB
$8 / 小时

▲NVIDIA DGX-2
2PFLOPS / 512GB
$400,000 / 台

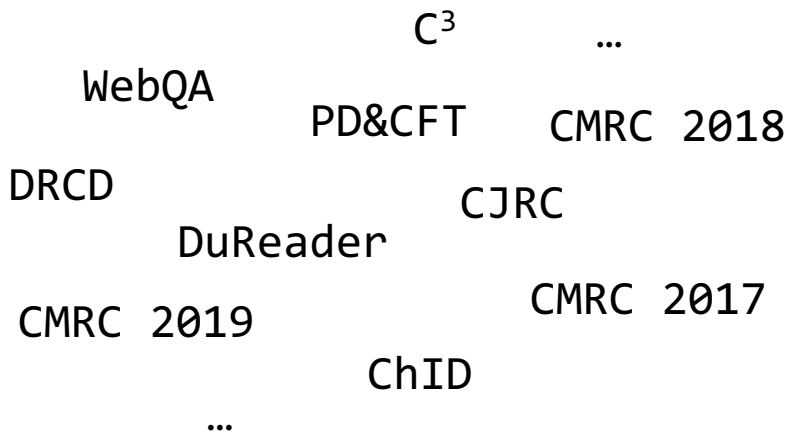
▲Google Cloud TPU v3 Pod
100PFLOPS / 32TB
$32 / 小时

技术发展热点（2）：跨语言

- **面向低资源语种的阅读理解研究**
  - 多数阅读理解研究及数据以英文为主，各个语种间的发展存在较大的差距
  - 如何利用英文数据或知识来帮助低资源语言的阅读理解效果成为一大问题



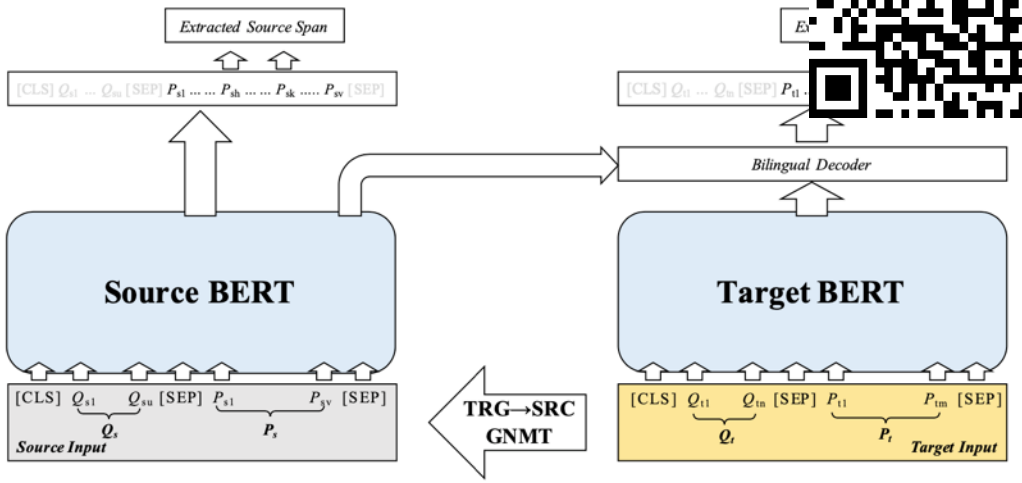
▲英文阅读理解数据集



▲中文阅读理解数据集

技术发展热点（2）：跨语言

- **Cross-Lingual Machine Reading Comprehension**
- **背景**：多数阅读理解研究集中于解决英文数据上的问题，其原因在于其他语种的阅读理解数据并不充足
- **目标**：构建一个跨语言的系统，能够有效利用富资源语种的语料来提升低资源语种的阅读理解系统效果
- **主要贡献**
  - 给出若干zero-shot方法，在中文（简体、繁体）、日语、法语阅读理解数据上获得显著性能提升
  - 提出Dual BERT双语编码器，对同一问题在双语环境中建模，从而丰富语义表示
  - 分析了在跨语言阅读理解任务中的重要因素，为未来研究提供了良好的基础



#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
	Human Performance	*91.1*	*97.3*	*92.4*	*97.9*	*90.4*	*95.2*	-	-	*80.4*	*93.3*
	P-Reader (single model)†	59.9	81.5	65.2	84.4	15.1	39.6	-	-	-	-
	Z-Reader (single model)†	79.8	92.7	74.2	88.1	13.9	37.4	-	-	-	-
	MCA-Reader (ensemble)†	66.7	85.5	71.2	88.1	15.5	37.1	-	-	-	-
	RCEN (ensemble)†	76.3	91.4	68.7	85.8	15.3	34.5	-	-	-	-
	r-net (single model)†	-	-	-	-	-	-	-	-	29.1	44.4
	DA (Yang et al., 2019)	49.2	65.4	-	-	-	-	55.4	67.7	-	-
1	GNMT+BERT$_{SQ-B_{en}}$ ♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$ ♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6
6	BERT$_{B_{cn}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	BERT$_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3
9	BERT$_{SQ-B_{mul}}$ ♠	56.5	77.5	59.7	79.9	18.6	41.4	66.7	81.0	65.4	80.1
10	BERT$_{SQ-B_{mul}}$ + Cascade Training	66.6	87.3	71.8	89.4	25.6	52.3	85.2	91.4	84.4	90.8
11	BERT$_{B_{mul}}$ + Mixed Training	66.8	87.5	72.6	89.8	26.7	53.4	85.3	91.6	84.7	91.2
12	**Dual BERT (w/ SQuAD)**	68.0	88.1	73.6	90.2	27.8	55.2	86.0	92.1	85.4	91.6

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. Cross-Lingual Machine Reading Comprehension. ***EMNLP 2019****.*

技术发展热点（3）：可解释

- **阅读理解过程的可解释性，甚至是人工智能技术的可解释性**
  - 知其然更要知其所以然，做到有据可循
  - 有时有必要牺牲部分“准确度”换取“可解释性”
  - 在一些领域仅仅给出最终答案并不能够信服用户，例如：司法、医疗领域
  - 在给出答案的同时能够给出相关“证据”，“线索”，“推理路径”

总结

- **机器在某些数据上已超过“人类水平”，但绝非代表机器已经具备足够的理解能力**
  - 机器在依靠“匹配”等简单模式的阅读理解任务中表现优异
  - 涉及多句推理的情况，机器并不能够给出满意的结果
- **问题思考**
  - 在“预训练模型”时代，是不是无脑堆数据就可以了？语言学分析还有没有必要？
  - 跨越语言的限制，如何能够利用英文资源来帮助低资源语言的系统效果？
- **未来发展**
  - 设计更加精妙的预训练方法以及预训练模型的压缩
  - 关注跨语言方法，通过“借力”的方式提升稀缺资源语种的系统性能
  - 探求“可解释”的阅读理解，为用户提供更为可靠的人工智能技术

哈工大讯飞联合实验室（HFL）

科大讯飞 iFLYTEK　HIT-SCIR

- **哈工大讯飞联合实验室成立于2014年9月，今年将迎来五周年**
  - 高校：哈尔滨工业大学社会计算与信息检索研究中心（HIT-SCIR）
  - 企业：科大讯飞AI研究院（iFLYTEK AI Research）
  - 重点研究方向：阅读理解、自动问答、人机对话、智能司法、智慧教育等
  - 双方在认知智能领域展开全面合作，重点突破语义理解、推理决策等关键技术

SQuAD评测冠军

SemEval 2018冠军

CGED评测冠军

CoQA评测冠军

SQuAD 2.0评测冠军

QuAC评测冠军

- **欢迎关注“哈工大讯飞联合实验室”微信公众号了解更多信息**

CSDN

AI ProCon 开发者大会

谢谢

ymcui@iflytek.com