CCL 2018自然语言处理国际前沿动态综述
阅读理解方向

崔一鸣

哈工大讯飞联合实验室(HFL)，科大讯飞

2018-10-21

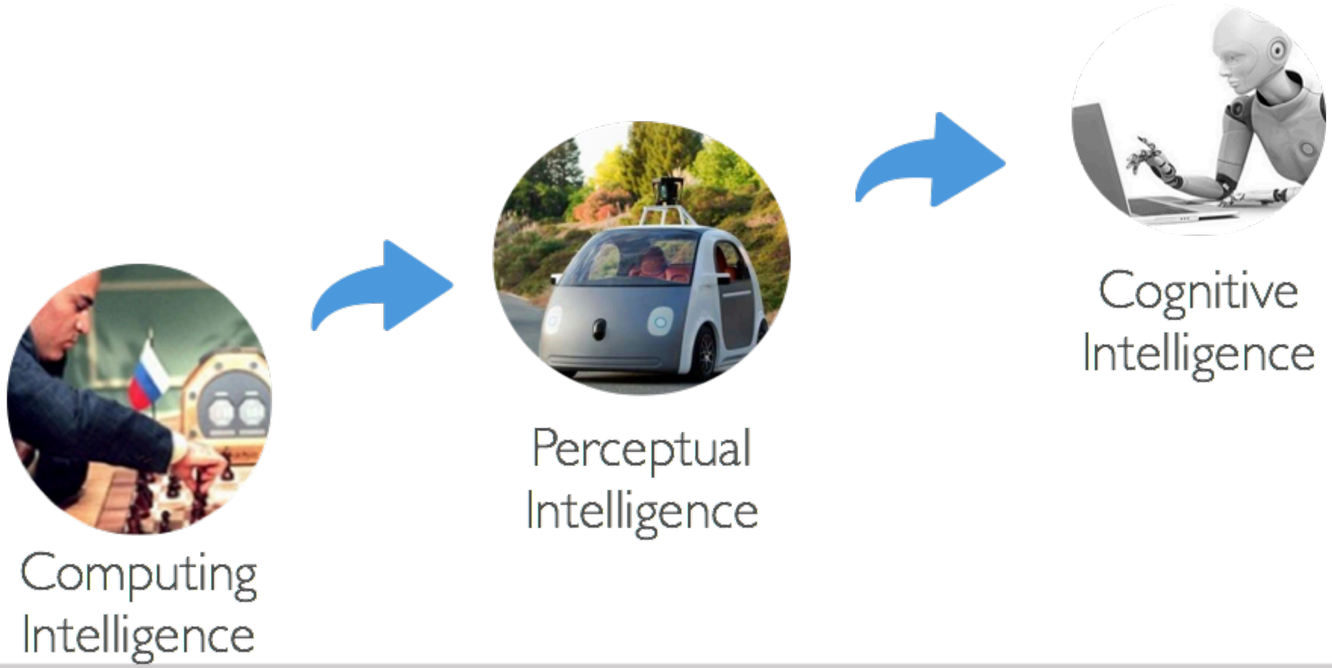
内容概要

- 阅读理解任务简介
- 阅读理解领域近期研究趋势
  - 数据集（任务）
  - 模型方法
  - 技术评测
- 报告总结
- 参考文献

阅读理解任务简介

阅读理解任务简介

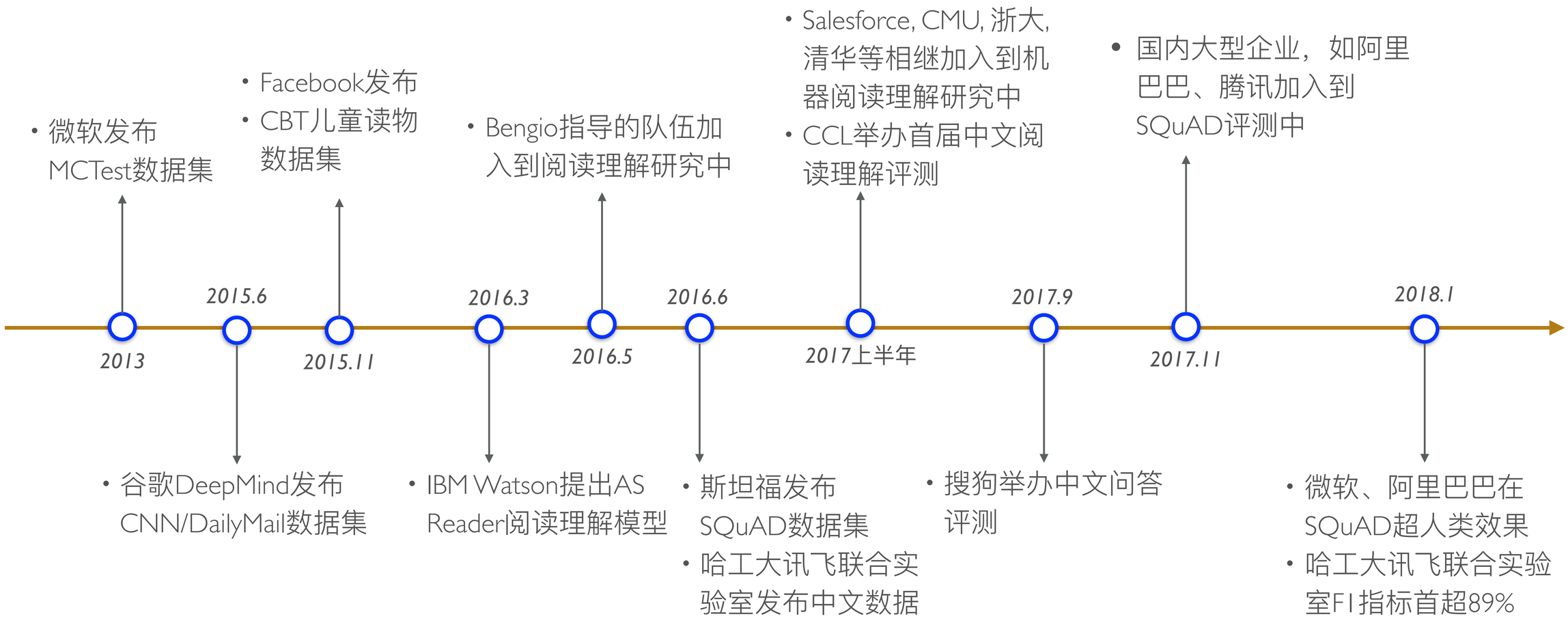
- 人工智能的一个重要目标是**让机器能听会说，能理解会思考**
- **机器阅读理解** (**M**achine **R**eading **C**omprehension, MRC) 作为认知智能的典型任务受到国内外研究人员的广泛关注



阅读理解任务简介

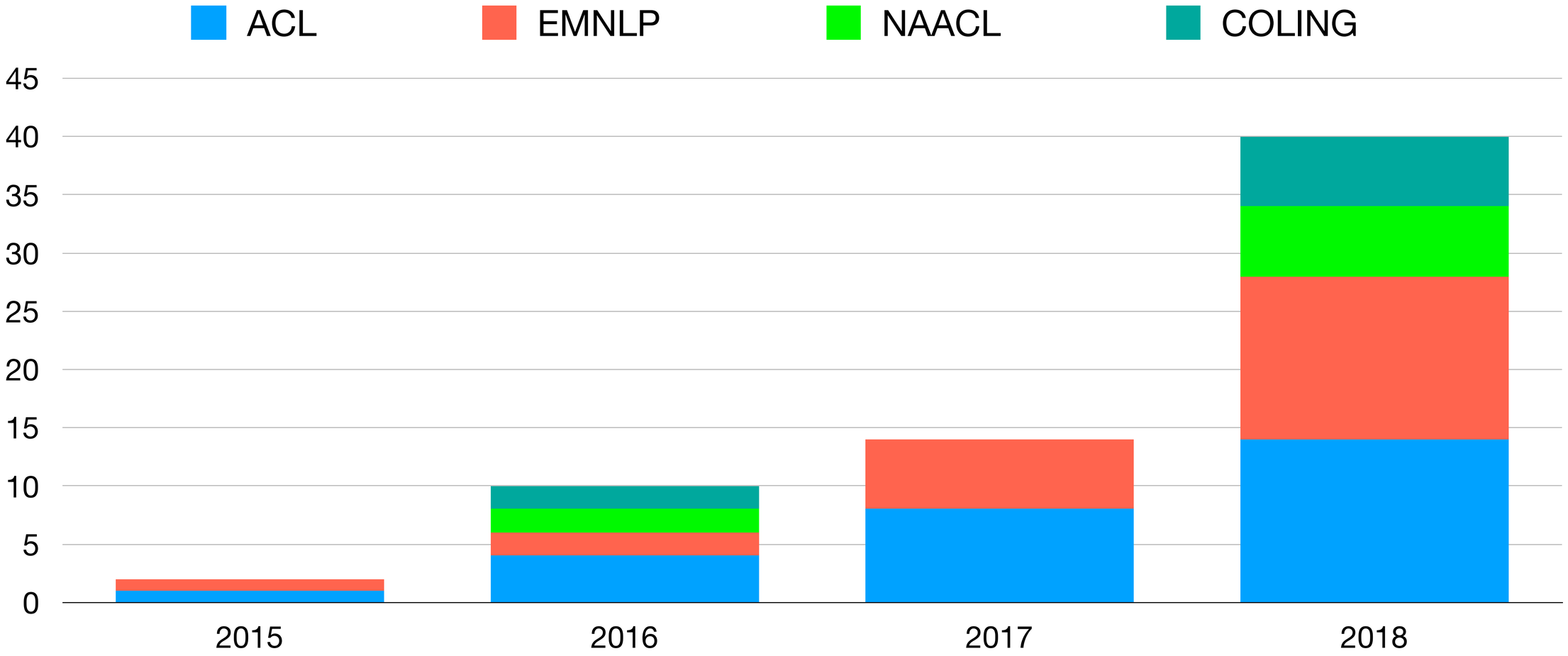
- 多数情况下，阅读理解任务包含如下几个要素
  - **D**ocument：需要机器阅读的篇章
    - 根据篇章的数量，可分为单文档阅读理解以及多文档阅读理解等
  - **Q**uestion：根据篇章内容所提出的问题
    - 根据问题的类型，可分为填空型或者用户提问型等
  - **C**andidate：候选答案
    - 根据任务的不同，会有一些候选答案，例如选择型阅读理解等
  - **A**nswer：答案
    - 根据任务的不同，答案可能是单个词、篇章片段、生成出来的句子等

前期阅读理解发展


2013
• 微软发布MCTest数据集
2015.6
• 谷歌DeepMind发布CNN/DailyMail数据集
2015.11
• Facebook发布
• CBT儿童读物数据集
2016.3
• IBM Watson提出AS Reader阅读理解模型
2016.5
• Bengio指导的队伍加入到阅读理解研究中
2016.6
• 斯坦福发布SQuAD数据集
• 哈工大讯飞联合实验室发布中文数据
2017上半年
• 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
2017.9
• 搜狗举办中文问答评测
2017.11
• 国内大型企业，如阿里巴巴、腾讯加入到SQuAD评测中
2018.1
• 微软、阿里巴巴在SQuAD超人类效果
• 哈工大讯飞联合实验室F1指标首超89%

阅读理解领域近期研究趋势

论文数量趋势

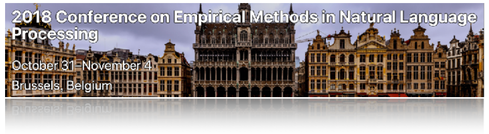


*相关数据统计自ACL Anthology: http://aclanthology.info/

**NAACL无17年数据

***COLING只在双数年份召开

论文数量趋势



*相关数据统计自ACL Anthology: http://aclanthology.info/

论文标题词云

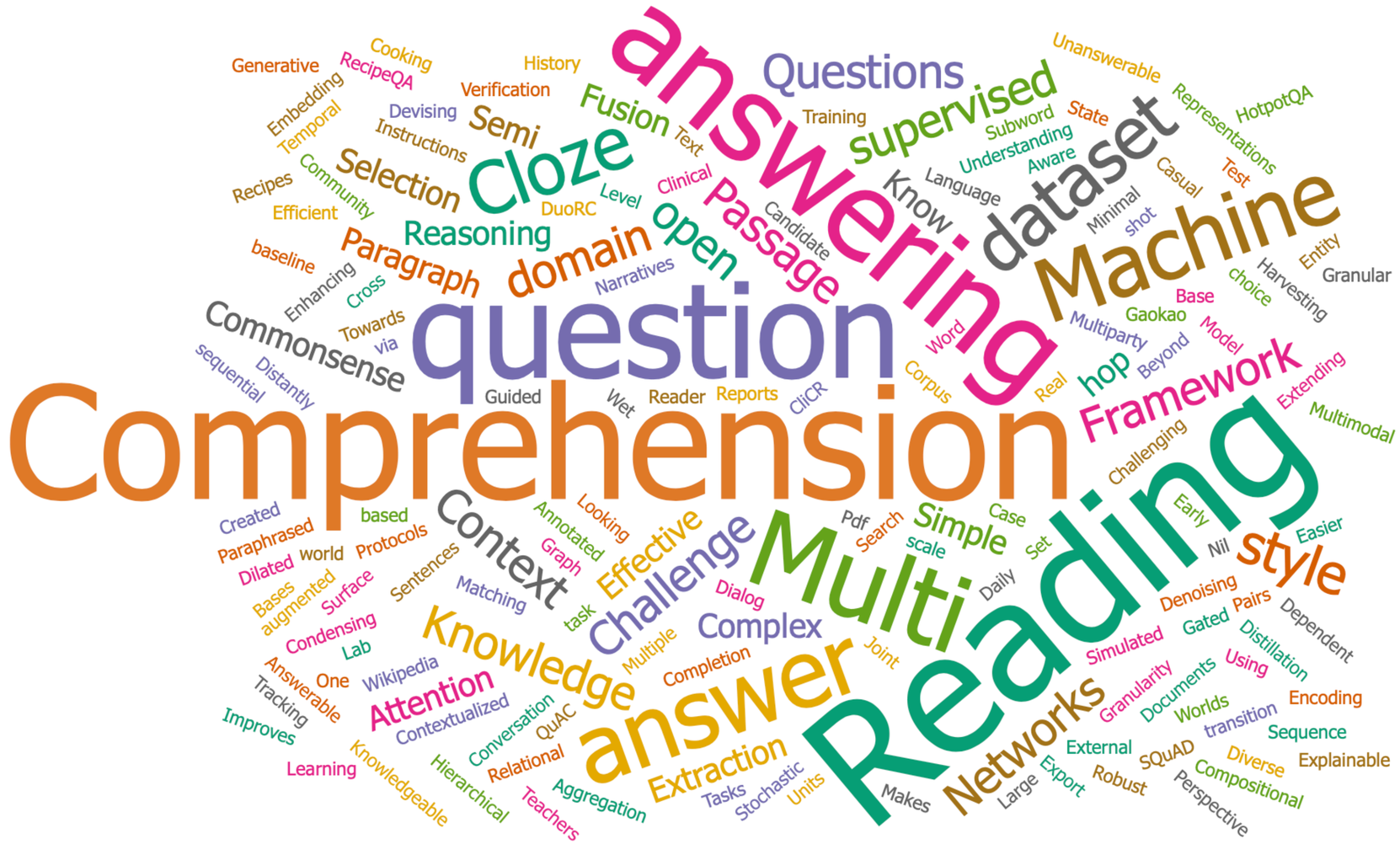
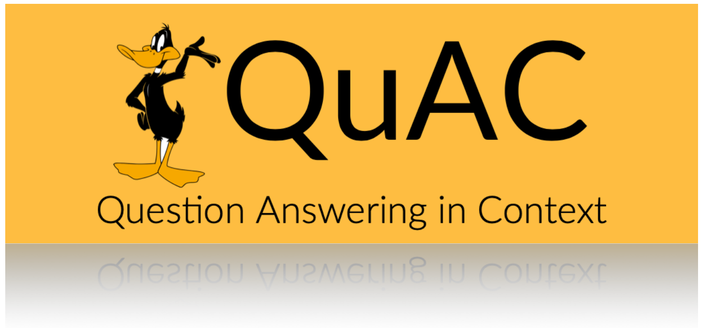


*Image generated by https://wordsift.org

数据集发展趋势

- 面向更复杂，更接近真实应用场景的数据集



SQuAD
The Stanford Question Answering Dataset

SQuAD2.0
The Stanford Question Answering Dataset

让世界聆听我们的声音

- **SQuAD 1.1 → SQuAD 2.0**
- 2016年6月，斯坦福大学发布了Stanford Question Answering Dataset (SQuAD) 数据集，开启了阅读理解数据集的新篇章
- 2018年7月，斯坦福大学发布了SQuAD 2.0，对原有数据集进行了扩充，加入了“**不可回答的问题**”，进一步考验机器阅读能力
  - 任务类型：预测问题是否可答，对于可答问题给出篇章中的连续片段作为答案
  - 数据量：10万个问题

Know What You Don't Know: Unanswerable Questions for SQuAD

Pranav Rajpurkar* **Robin Jia*** **Percy Liang**
Computer Science Department, Stanford University
{pranavsr,robinjia,pliang}@cs.stanford.edu

Article: Endangered Species Act
Paragraph: "*. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

科大讯飞 iFLYTEK

CoQA

A Conversational Question Answering Challenge

- **CoQA**: **Co**nversational **Q**uestion **A**nswering
- 阅读理解任务向其他NLP任务的渗透
- 斯坦福大学提出了CoQA数据集用于评测阅读理解技术在**多轮对话**过程中的应用
  - 任务类型：两个人阅读给定的一段短文，一个人来提问，另一个人来回答
  - 数据量：12.7万个问题
- **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**问题对问题理解提出新的挑战

CoQA: A Conversational Question Answering Challenge

Siva Reddy **Danqi Chen** **Christopher D. Manning**
Computer Science Department
Stanford University
{sivar,danqi,manning}@cs.stanford.edu

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

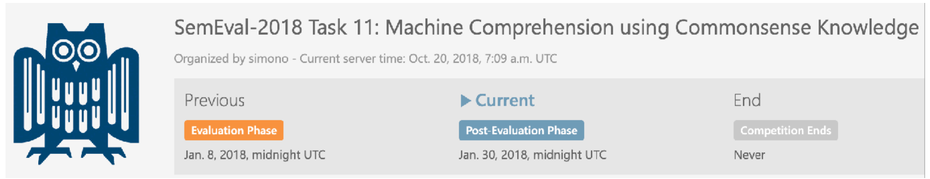
Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

技术发展趋势-I

- **第一点：重视引入外部知识**
- **当前问题**
  - 多数阅读理解模型采用完全端到端的模型结构，很难了解其中的运行机理
  - 某些问题除了篇章本身的信息之外还需要外部知识才能正确解答
- **解决方案**
  - 建立相关数据集，使得问题需要融合外部知识才能解答
  - 在现有模型基础上运用外部知识，进一步提升阅读理解系统效果
- 代表工作：SemEval-2018 Task 11, ARC Challenge等

技术发展趋势-I

- **SemEval-2018 Task 11: Machine Comprehension using Commonsense Knowledge**
  - 任务描述：对给定的篇章进行理解，并根据给定问题从两个候选答案中挑选出一个正确答案（其中回答问题会涉及到常识）
  - 考查内容：上下文理解、常识判断
- **AI2 ARC Challenge**
  - 形式：从自由文本中抽取知识，解答给定的单项选择题
  - 包括：14M自由文本知识，7787个选择题
  - 目前最优系统在挑战集效果上仅能达到约30%的准确率（随机猜测约25%）



技术发展趋势-II

- **第二点：采用大规模数据预训练的模型**
- **当前问题**
  - 多数阅读理解数据集的训练样本较少（几千 ~ 十几万）
  - 阅读理解更多依赖的是上下文的理解，而非对某一个单词的关注，独立的词向量很难表征更加丰富的语义信息
- **解决方案**
  - 采用大规模数据预训练的模型一定程度上缓解数据稀疏的问题（OOV等）
  - 根据上下文动态生成文本的表示，在向量表示中包含更多文本语义层次信息
- 主要代表工作：ELMo、OpenAI GPT、BERT等

科大讯飞 iFLYTEK

技术发展趋势-II

- **ELMo**: **E**mbeddings from **L**anguage **Mo**dels
  - NAACL-HLT-2018最佳论文奖，发布至今引用量已超过120+
  - 采用大规模语料（1B），使用大隐层维度（输出维度1024）训练双向语言模型BiLM
  - 训练使用3张GTX 1080，训练10轮共计花费两周时间
  - 简单易用，直接加在词向量平行的位置中
  - 目前已成为（在计算资源充足情况下）各NLP任务的标配

Deep contextualized word representations

Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

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%

Peters et al., 2018. Deep Contextualized Word Representations

技术发展趋势-II

- **BERT**: **B**idirectional **E**ncoder **R**epresentation from **T**ransformers
  - 在Open AI Transformers的基础上进一步提出深层双向Transformers
  - 无需设计复杂的任务相关的网络设计
  - 训练使用了BookCorpus以及Wikipedia和数据，共计词数量达到了33B
  - 训练使用了16*4=64块**TPU**，共计**4天时间**
    - 有人估算过使用GPU需要将近**1年的时间**
  - 官方宣称10月末放出相关预训练模型和代码

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

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**

Devlin et al., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

技术发展趋势-III

- **第三点：阅读理解与其他NLP任务的结合**
- 解决阅读理解任务的主要要点可归纳为三点
  1. 对于篇章内容的理解
  2. 对于问题的理解
  3. 篇章和问题之间的联系
- **篇章和问题的理解**是阅读理解任务中最基础且不可或缺的内容
- 篇章和问题的理解（狭义） → 上下文理解（广义）

技术发展趋势-III

- 阅读理解技术应用在多轮对话生成任务
  - 篇章和问题的理解（阅读理解） → 对话流和Query的理解（多轮对话）
  - 在Ubuntu、OpenSubtitles等数据集上获得良好效果的

Context-Sensitive Generation of Open-Domain Conversational Responses

Wei-Nan Zhang*, Yiming Cui†, Yifa Wang*, Qingfu Zhu*, Lingzhi Li*, Lianqiang Zhou‡, Ting Liu*
*Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China.
†Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
‡Joint Laboratory of HIT and Tencent Corporation, Shenzhen, China
*{wnzhang, yfwang, qfzhu, lzli, tliu}@ir.hit.edu.cn
†ymcui@iflytek.com
‡tomcatzhou@tencent.com

Models	Ubuntu			OpenSubtitles		
	Coherence	Naturalness	Diversity	Coherence	Naturalness	Diversity
LSTM	0.930	0.477	0.069	0.963	0.443	0.099
HRED	0.967	0.490	0.141	0.963	0.443	0.098
VHRED	1.010	0.507	0.140	0.986	0.473	0.093
CVAE	0.987	**0.513**	0.140	1.000	0.477	**0.114**
WSI	1.010	0.507	0.141	1.013	0.490	0.110
HRAN	1.027	0.510	0.147	**1.033**	0.477	0.109
Dynamic	0.987	0.507	**0.158**	1.013	0.477	0.109
Static	**1.070**	**0.513**	0.150	1.027	**0.497**	0.110

Zhang et al., COLING2018. Context-sensitive Generation of Open-Domain Conversational Responses

技术发展趋势-III

- **第四点：面向特定数据集的研究**
- SQuAD 1.1：仅预测篇章片段
  - 设计更加复杂的篇章和问题的交互计算：SAN (Liu et al., 2018)，SLQA (Wang et al., 2018)
  - 生成额外的训练数据：QANet (Yu et al., 2018),
- SQuAD 2.0：判断答案是否可答，对于可答问题需要预测篇章片段
  - 设计额外的损失函数：RMR+Verifier (Hu et al., 2018), U-Net (Sun et al., 2018)
  - 答案验证：RMR+Verifier, Answer Verifier (Tan et al., 2018)
- 选择型阅读理解：从给定的多个选项中选出一个正确答案
  - 计算篇章、问题、选项三者之间的关系：Co-Matching (Wang et al., 2018), HMA (Chen et al., 2018)

中文技术评测

- 各大研究机构纷纷举办阅读理解相关技术评测，中文机器阅读理解技术研究持续推进，阅读理解技术的受众语种范围不断扩大
- CCL 2017**首次举办**了中文机器阅读理解评测，今年是系列评测的第二届
  - 主办方：中国中文信息学会计算语言学专委会（CIPS-CL）
  - 承办方：哈工大讯飞联合实验室（HFL）
  - 冠名方：科大讯飞股份有限公司
- 两届评测共吸引了超过200支队伍报名，报名人数近千人，形成了良好口碑

报告总结

报告总结

- 数据集：面向更接近真实应用场景，难度更大的阅读理解数据集
- 技术趋势
  - 对于特定类型问题引入外部知识加以解决
  - 大规模数据预训练模型的应用
  - 交叉任务之间的联系，将阅读理解技术推广到其他NLP任务中
- 中文阅读理解研究蓬勃发展，研究队伍不断壮大
- 关注效果提升，但也更要关注模型没能解决的问题

参考文献

- ACL Anthology: http://aclanthology.info
- Rajpurkar et al., 2018. Know What You Don’t Know: Unanswerable Questions for SQuAD.
- Choi et al., 2018. QuAC: Question Answering in Context.
- Yagcioglu et al., 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes.
- Yang et al., 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering.
- Reddy et al., 2018. CoQA: A Conversational Question Answering Challenge.
- Rajpurkar et al., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text.
- Clark et al., 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.
- Peters et al., 2018. Deep contextualized word representations.
- Devlin et al., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

参考文献

- Liu et al,. 2018. Stochastic Answer Networks for Machine Reading Comprehension.
- Wang et al., 2018. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering.
- Yu et al., 2017. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.
- Hu et al., 2018. Read + Verify: Machine Reading Comprehension with Unanswerable Questions.
- Sun et al., 2018. U-Net: Machine Reading Comprehension with Unanswerable Questions.
- Tan et al., 2018. I Know There Is No Answer: Modeling Answer Validation for Machine Reading Comprehension.
- Wang et al., 2018. A Co-Matching Model for Multi-choice Reading Comprehension.
- Chen et al., 2018. HFL-RC System at SemEval-2018 Task 11: Hybrid Multi-Aspects Model for Commonsense Reading Comprehension.
- Zhang et al., 2018. Context-sensitive Generation of Open-Domain Conversational Responses.

谢谢大家