机器阅读理解和对话系统

张倬胜

上海交通大学

zhangzs@sjtu.edu.cn

http://bcmi.sjtu.edu.cn/~zhangzs

2021/07/29

目录

个人简介

张倬胜

博士研究生，上海交通大学计算机科学与工程系

指导老师：赵海教授

研究兴趣：语言模型、阅读理解、对话系统

个人主页：http://bcmi.sjtu.edu.cn/~zhangzs

教育/经历：

- **2020.9-至今** **上海交通大学** 计算机科学与技术 **博士（在读）**
- **2019.06-2020.07** **NICT (Japan)** | Internship Research Fellow
- **2016.09-2020.03** **上海交通大学** 计算机科学与技术 **获硕士学位**
- **2016.06-2016.09** **IBM Watson Team** | Data Scientist Intern
- **2012.09-2016.06** **武汉大学** 计算机学院 物联网工程 **获学士学位**

主要荣誉：

- **全球AI华人百强学术新星**
- 阅读理解榜单与评测**第一名**：SQuAD2.0、RACE、ShARC、MuTual、CMRC、SNLI等
- 上海交通大学研究生学术之星
- CCF优秀大学生

排行榜：阅读理解CMRC 2017

- 中文机器阅读理解大赛CMRC2017 **第一名（最佳单系统）**

最佳单系统（Best Single System）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	单系统	76.15%	77.73%

最终系统排名

填空类问题（Cloze-style Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	6ESTATES PTE LTD	多系统	81.85%	81.90%
		单系统	75.85%	74.73%
2	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	多系统	78.35%	80.67%
		单系统	76.15%	77.73%
3	南京云思创智信息科技有限公司	多系统	79.20%	80.27%
		单系统	77.15%	77.53%
4	华东师范大学 East China Normal University (ECNU)	多系统	79.45%	79.70%
		单系统	77.95%	77.40%
5	鲁东大学 Ludong University	多系统	77.05%	77.07%
		单系统	74.75%	75.07%
6	武汉大学语言与信息研究中心 Wuhan University (WHU)	单系统	78.20%	76.53%

排行榜：自然语言推理SNLI (2018-2019)

- 斯坦福自然语言推理(snli)排行榜 **第一名**

Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1

排行榜：阅读理解SQuAD2.0挑战赛 (2019-2020)

- **Stanford大学**提出的排行榜竞赛，已成为机器阅读理解**顶级赛事**
- 2019年：**首次**以单模型超越**人类基准**，并获得**第一名**
- 2020年：单模型和混合模型均获得**第一名**

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071
5 Jul 26, 2019	UPM (single model) *Anonymous*	87.193	89.934
6 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
6 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795

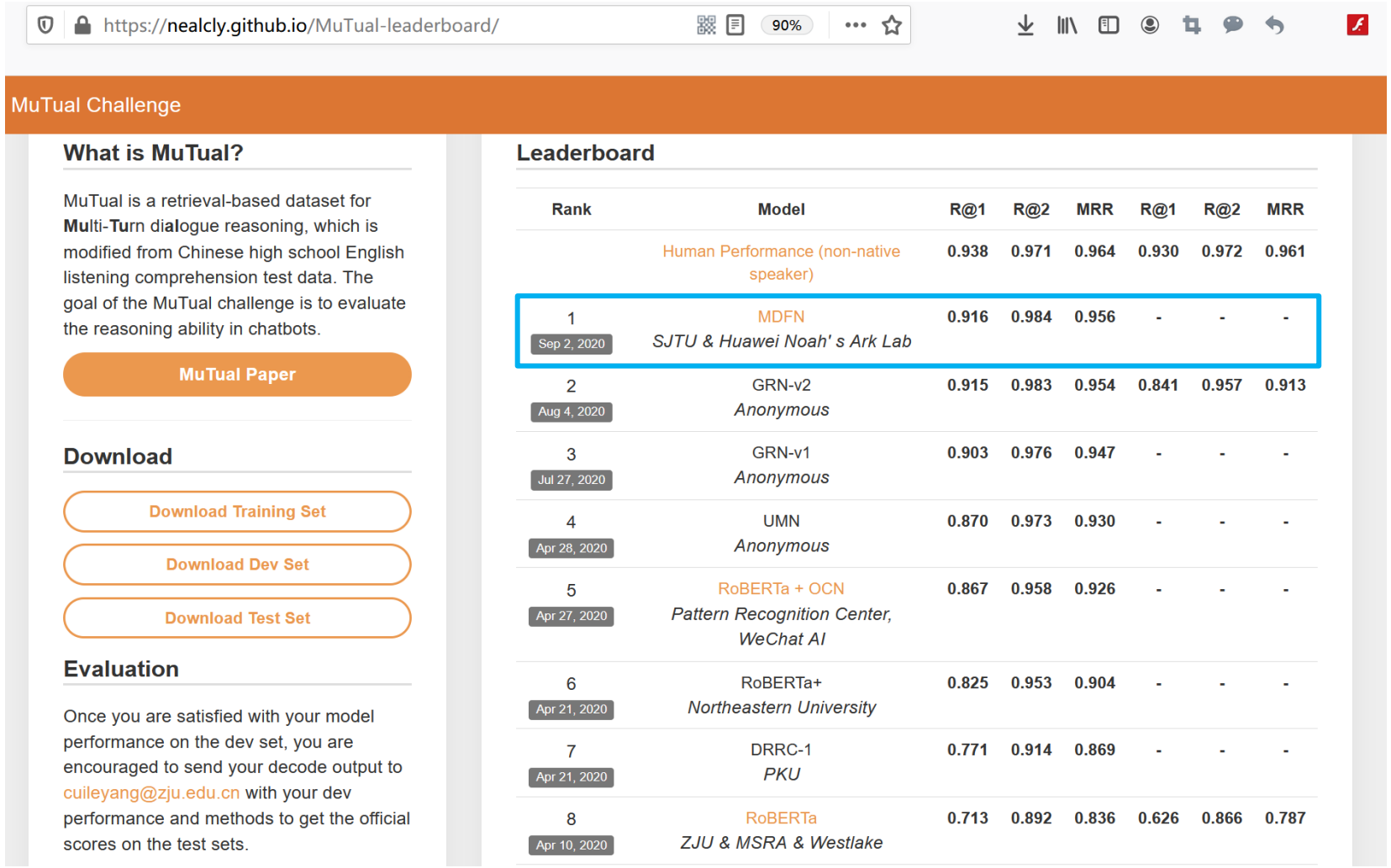
SQuAD2.0 排行榜 2019.07-09

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	90.115	92.580
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
5 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019

SQuAD2.0 排行榜 2020.01-

排行榜：多轮对话推理 MuTual (2020)

- 多轮对话推理排行榜MuTual **第一名**



- 题目类型：高中英语听力题

排行榜：会话式机器阅读理解ShARC (2021)

- 会话式机器阅读理解ShARC**第一名**

ShARC: End-to-end Task

#	Model / Reference	Affiliation	Date	Micro Accuracy[%]	Macro Accuracy[%]	BLEU-1	BLEU-4
1	DGM	Shanghai Jiao Tong University	Jan 2021	**77.4**	**81.2**	63.3	48.4
2	Discern (single model)	The Chinese University of Hong Kong	May 2020	73.2	78.3	**64.0**	49.1
3	EMT	Salesforce Research & CUHK	Nov 2019	69.4	74.8	60.9	46.0
4	EMT + entailment	Salesforce Research & CUHK	Mar 2020	69.1	74.6	63.9	**49.5**
5	UrcaNet (ensemble)	IBM Research AI	Dec 2019	69.0	74.6	56.7	42.0
6	E3	University of Washington	Feb 2019	67.6	73.3	54.1	38.7
7	BiSon (single model)	NEC Laboratories Europe	Aug 2019	66.9	71.6	58.8	44.3
8	UrcaNet (single model)	IBM Research AI	Aug 2019	65.1	71.2	60.5	46.1

- 包含“对话决策”和“问题生成”两个子任务

Rule Text: *Eligible applicants may obtain direct loans for up to a maximum indebtedness of $300,000, and guaranteed loans for up to a maximum indebtedness of $1,392,000 (amount adjusted annually for inflation).*

User Scenario: *I got my loan last year. It was for 450,000.*

Initial Question: *Does this loan meet my needs?*

Decision: Yes | No | **Inquire** | Irrelevant

Follow-up Q1: *Do you need a direct loan?*

Follow-up A1: *Yes.*

Decision: Yes | No | **Inquire** | Irrelevant

Follow-up Q2: *Is your loan for less than 300,000?*

Follow-up A2: *No.*

Decision: Yes | No | **Inquire** | Irrelevant

Follow-up Q3: *Is your loan less than 1,392,000?*

Follow-up A2: *Yes.*

Decision: **Yes** | No | Inquire | Irrelevant

Final Answer: *Yes.*

目录

机器阅读理解

- 目标: 让机器理解人类语言并解决现实问题
- 挑战：问题表述多样性、语言歧义、图谱知识有限、缺乏推理能力、不懂常识等等
- 主流形式：中高考阅读理解题型（选择、填空、问答等）
- 阅读理解任务从2015年开始逐渐得到广泛关注
  - 训练机器阅读文本和学习知识，解决语义理解问题
  - 阅读长文本文章，对相应的问题进行解答
  - 现实应用：自动问答系统、对话机器人、金融分析、医疗知识理解等

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

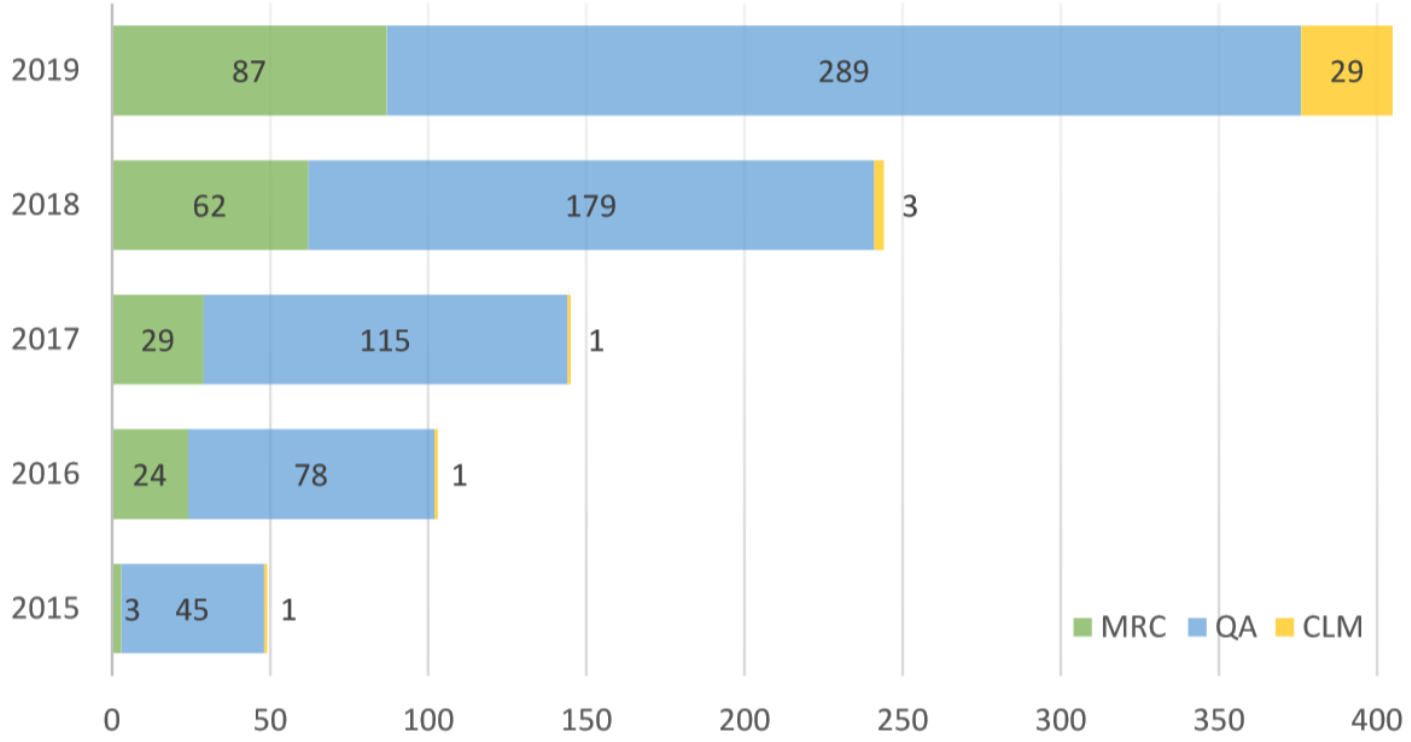
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

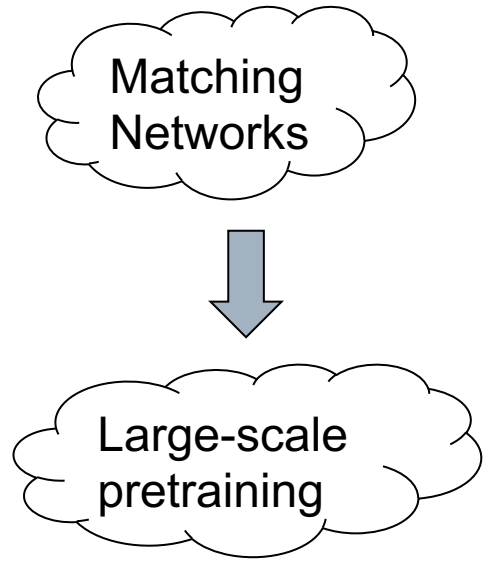
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

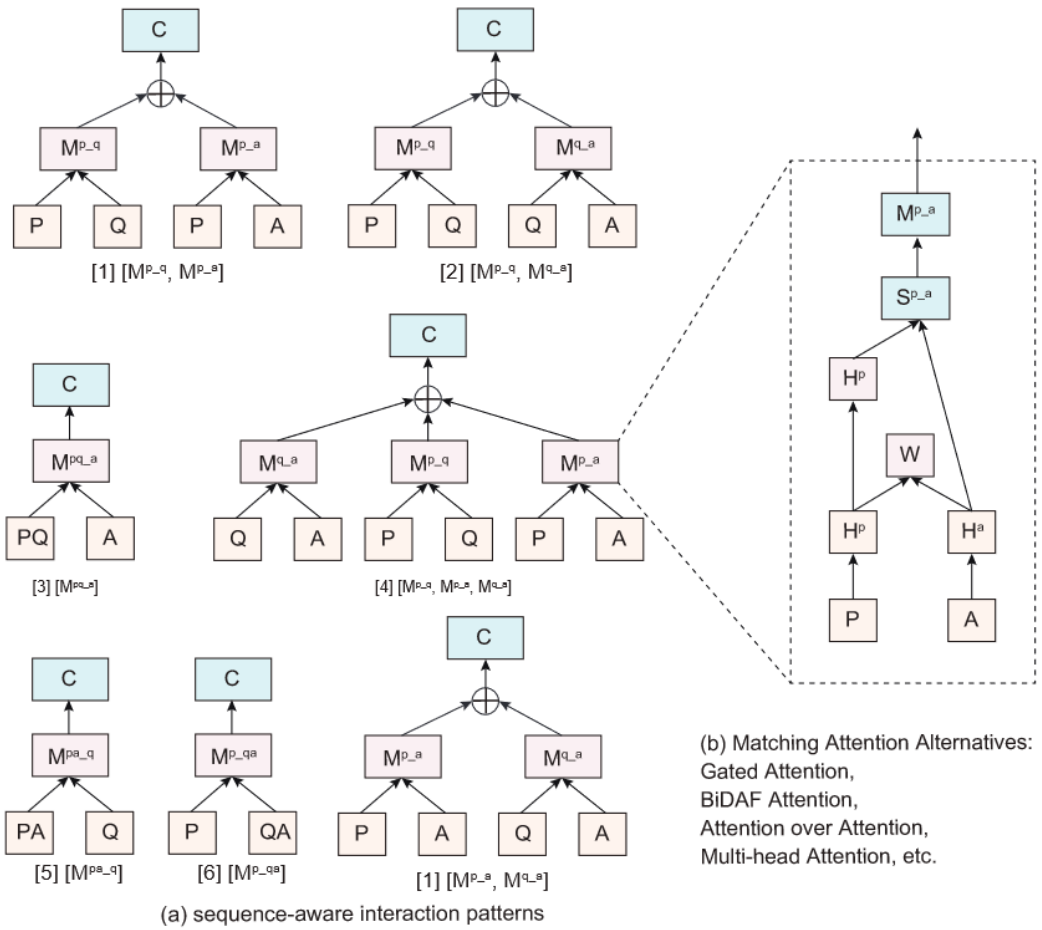
机器阅读理解

- 阅读理解研究经历两大热潮
  - 深度神经网络，尤其是注意力机制的广泛应用
  - 预训练语言模型大大提高基准水平





第一阶段: 匹配网络



Method	Att. Type	CNN val	CNN test	DailyMail val	DailyMail test
Attentive Reader (Hermann et al. 2015)	UA	61.6	63.0	70.5	69.0
AS Reader (Kadlec et al. 2016)	UA	68.6	69.5	75.0	73.9
Iterative Attention (Sordoni et al. 2016)	UA	72.6	73.3	-	-
Stanford AR (Chen, Bolton, and Manning 2016)	UA	73.8	73.6	77.6	76.6
GAReader (Dhingra et al. 2017)	UA	73.0	73.8	76.7	75.7
AoA Reader (Cui et al. 2017)	BA	73.1	74.4	-	-
BiDAF (Seo et al. 2017)	BA	76.3	76.9	80.3	79.6

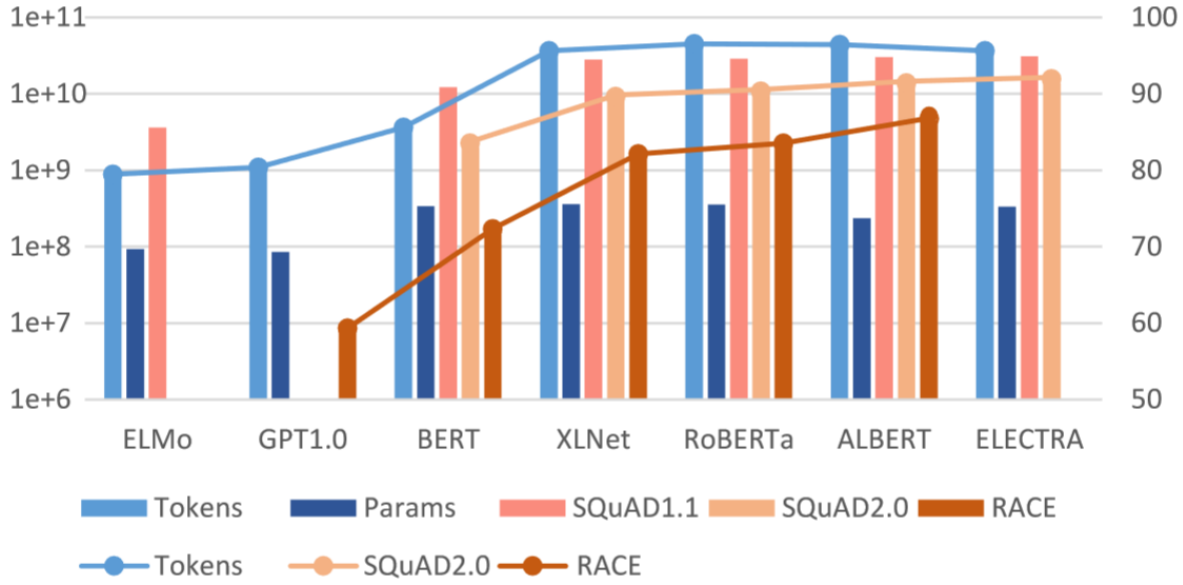
Model	Matching	M	H	RACE
Human Ceiling Performance (Lai et al. 2017)		95.4	94.2	94.5
Amazon Mechanical Turker (Lai et al. 2017)		85.1	69.4	73.3
HAF (Zhu et al. 2018a)	$[M^{P\_A}; M^{P\_Q}; M^{Q\_A}]$	45.0	46.4	46.0
MRU (Tay, Tuan, and Hui 2018)	$[M^{P\_Q\_A}]$	57.7	47.4	50.4
HCM (Wang et al. 2018a)	$[M^{P\_Q}; M^{P\_A}]$	55.8	48.2	50.4
MMN (Tang, Cai, and Zhuo 2019)	$[M^{Q\_A}; M^{A\_Q}; M^{P\_Q}; M^{P\_A}]$	61.1	52.2	54.7
GPT (Radford et al. 2018)	$[M^{P\_Q\_A}]$	62.9	57.4	59.0
RSM (Sun et al. 2019b)	$[M^{P\_QA}]$	69.2	61.5	63.8
DCMN (Zhang et al. 2019a)	$[M^{PQ\_A}]$	77.6	70.1	72.3
OCN (Ran et al. 2019a)	$[M^{P\_Q\_A}]$	76.7	69.6	71.7
BERT$_{large}$ (Pan et al. 2019b)	$[M^{P\_Q\_A}]$	76.6	70.1	72.0
XLNet (Yang et al. 2019c)	$[M^{P\_Q\_A}]$	85.5	80.2	81.8
+ DCMN+ (Zhang et al. 2020a)	$[M^{P\_Q}; M^{P\_O}; M^{Q\_O}]$	86.5	81.3	82.8
RoBERTa (Liu et al. 2019c)	$[M^{P\_Q\_A}]$	86.5	81.8	83.2
+ MMM (Jin et al. 2019a)	$[M^{P\_Q\_A}]$	89.1	83.3	85.0
ALBERT (Jin et al. 2019a)	$[M^{P\_Q\_A}]$	89.0	85.5	86.5
+ DUMA (Zhu, Zhao, and Li 2020)	$[M^{P\_QA}; M^{QA\_P}]$	90.9	86.7	88.0
Megatron-BERT (Shoeybi et al. 2019)	$[M^{P\_Q\_A}]$	91.8	88.6	89.5

第二阶段：预训练语言模型

Models	Encoder	EM	F1	↑ EM	↑ F1
Human (Rajpurkar, Jia, and Liang 2018)	-	82.304	91.221	-	-
Match-LSTM (Wang and Jiang 2016)	RNN	*64.744*	*73.743*	-	-
DCN (Xiong, Zhong, and Socher 2016)	RNN	66.233	75.896	1.489	2.153
Bi-DAF (Seo et al. 2017)	RNN	67.974	77.323	3.230	3.580
Mnemonic Reader (Hu, Peng, and Qiu 2017)	RNN	70.995	80.146	6.251	6.403
Document Reader (Chen et al. 2017)	RNN	70.733	79.353	5.989	5.610
DCN+ (Xiong, Zhong, and Socher 2017)	RNN	75.087	83.081	10.343	9.338
r-net (Wang et al. 2017)	RNN	76.461	84.265	11.717	10.522
MEMEN (Pan et al. 2017)	RNN	78.234	85.344	13.490	11.601
QANet (Yu et al. 2018)*	TRFM	80.929	87.773	16.185	14.030
CLMs					
ELMo (Peters et al. 2018)	RNN	78.580	85.833	13.836	12.090
BERT (Devlin et al. 2018)*	TRFM	85.083	91.835	20.339	18.092
SpanBERT (Joshi et al. 2020)	TRFM	88.839	94.635	24.095	20.892
XLNet (Yang et al. 2019c)	TRFM-XL	89.898	95.080	25.154	21.337

Models	Encoder	SQuAD 2.0	↑ F1	RACE	↑ Acc
Human (Rajpurkar, Jia, and Liang 2018)	-	91.221	-	-	
GPT_{v1} (Radford et al. 2018)	TRFM	-	-	*59.0*	-
BERT (Devlin et al. 2018)	TRFM	*83.061*	-	72.0	-
SemBERT (Zhang et al. 2020b)	TRFM	87.864	4.803	-	-
SG-Net (Zhang et al. 2020c)	TRFM	87.926	4.865	-	-
RoBERTa (Liu et al. 2019c)	TRFM	89.795	6.734	83.2	24.2
ALBERT (Lan et al. 2019)	TRFM	90.902	7.841	86.5	27.5
XLNet (Yang et al. 2019c)	TRFM-XL	90.689	7.628	81.8	22.8
ELECTRA (Clark et al. 2019c)	TRFM	91.365	8.304	-	-

Method	Tokens	Size	Params	SQuAD1.1 Dev	SQuAD1.1 Test	SQuAD2.0 Dev	SQuAD2.0 Test	RACE
ELMo	800M	-	93.6M	85.6	85.8	-	-	-
GPT_{v1}	985M	-	85M	-	-	-	-	59.0
$XLNet_{large}$	33B	-	360M	94.5	95.1*	88.8	89.1*	81.8
$BERT_{large}$	3.3B	13GB	340M	91.1	91.8*	81.9	83.0	72.0†
$RoBERTa_{large}$	-	160GB	355M	94.6	-	89.4	89.8	83.2
$ALBERT_{xxlarge}$	-	157GB	235M	94.8	-	90.2	90.9	86.5
$ELECTRA_{large}$	33B	-	335M	94.9	-	90.6	91.4	-



阅读理解同样重要

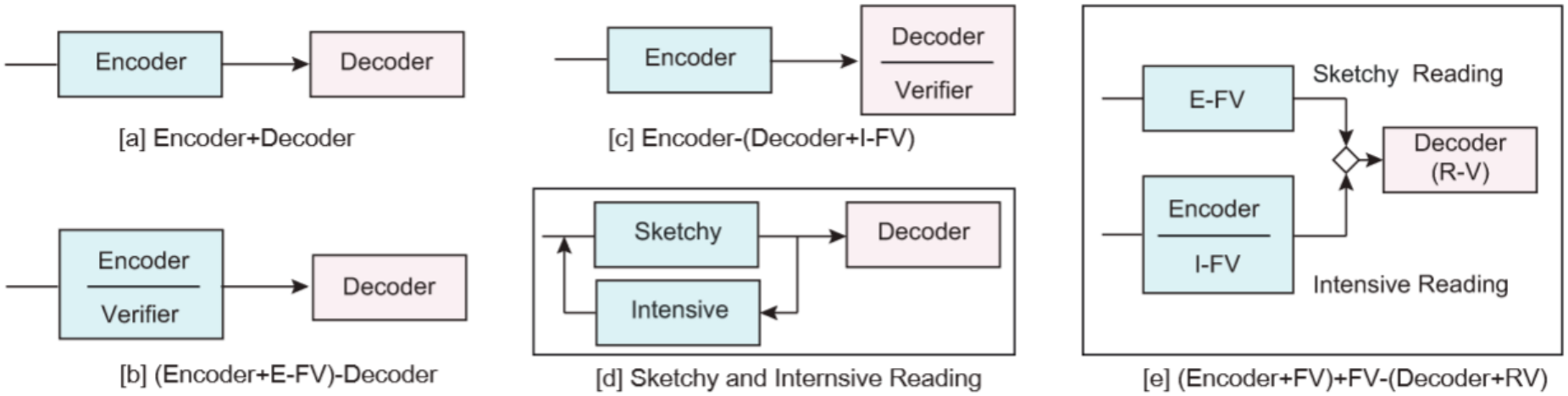
Reading Strategy based on human reading patterns

- Learning to skim text
- Learning to stop reading
- Retrospective reading
- Back and forth reading, highlighting, and self-assessment

Tactic Optimization:

- The **objective** of answer verification
- The **dependency** inside answer span
- **Re-ranking** of candidate answers

Answer Verification Design (From Retro-Reader)



目录

研究路线

2016年

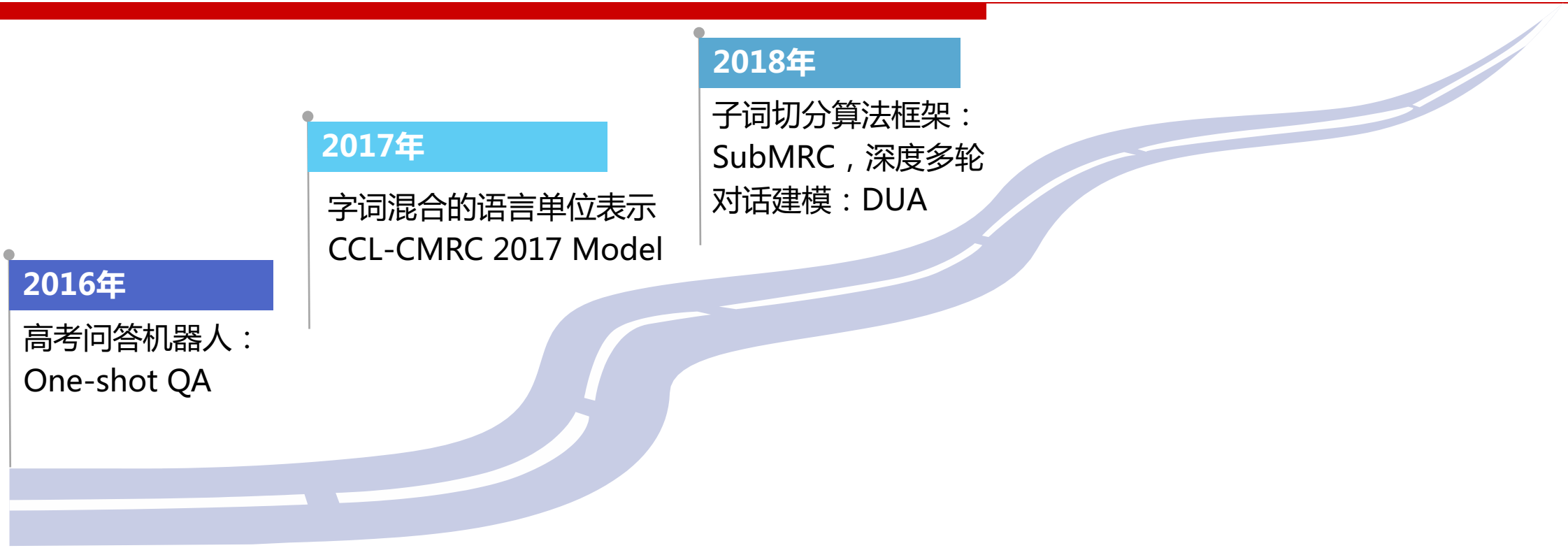
高考问答机器人：
One-shot QA

- **2016年 初探：高考机器人（历史）**
  - 目标：让机器人参加高考
  - 面临的问题：数据？模型？评估？
  - 做过的尝试+推倒重来：数据爬取规范化，自动构建知识图谱，少样本学习，排序评估方式

研究路线

- **2017年 破局：机器阅读理解**
  - 从解决实际任务入手：参加CMRC 2017阅读理解评测
  - 针对未登录词问题，提出了有效的字词融合建模和基于词频的平滑过滤机制
  - 获得了单系统第一名

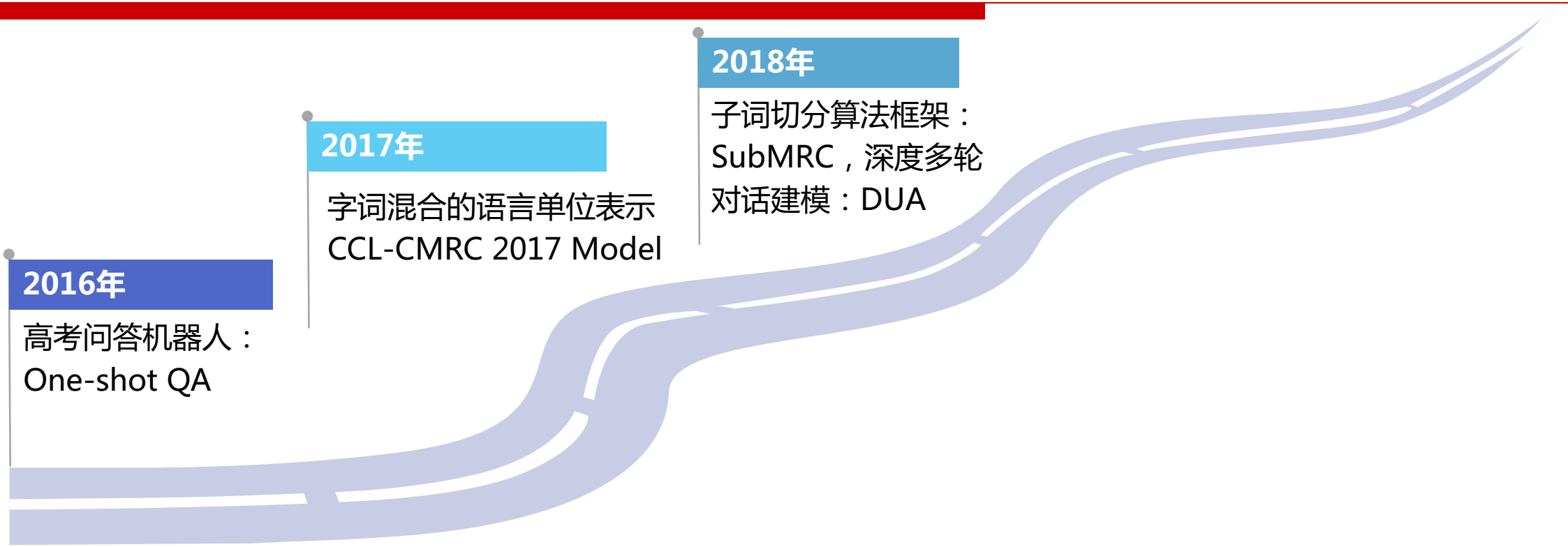
研究路线



- **2018年 进阶：更深入的语言理解研究**
  - 方法探索：寻找更灵活的语言单位粒度切分机制 (SubMRC, TASLP)
  - 任务延申：探索多轮对话理解任务中的对话选择匹配问题 (DUA，COLING 2018)

One-shot Learning for Question-Answering in Gaokao History Challenge
Subword-augmented Embedding for Cloze Reading Comprehension
Modeling Multi-turn Conversation with Deep Utterance Aggregation

研究路线



- **2019年 升华：从计算语言学的角度研究语言理解**
  - 针对词表大小限制以及词表固定的问题，提出使用动态词表进行语言建模 (OpenIME, ACL 2019)
  - 针对阅读理解中的语言理解鸿沟，使用语义角色信息指导语言建模 (SemBERT, AAAI 2020)
  - 针对阅读理解长文本依赖问题，提出使用句法引导Transformer中的注意力学习 (SG-Net , AAAI 2020)
  - 针对模态单一的问题，提出结合图像检索与文本表示合二为一的多模态语言表征 (UVR, ICLR 2020)

研究路线

2016年

高考问答机器人：
One-shot QA

2017年

字词混合的语言单位表示
CCL-CMRC 2017 Model

2018年

子词切分算法框架：
SubMRC，深度多轮
对话建模：DUA

2019年

词表自适应语言建模
语义感知模型：SemBERT
句法引导阅读理解系统：SG-Net
多模态语言表示模型：UVR

2020年

基于常识的对话理解：RekNet
任务导向的预训练：DAPO
多说话人层次化对话建模：MDFN
预训练语言模型

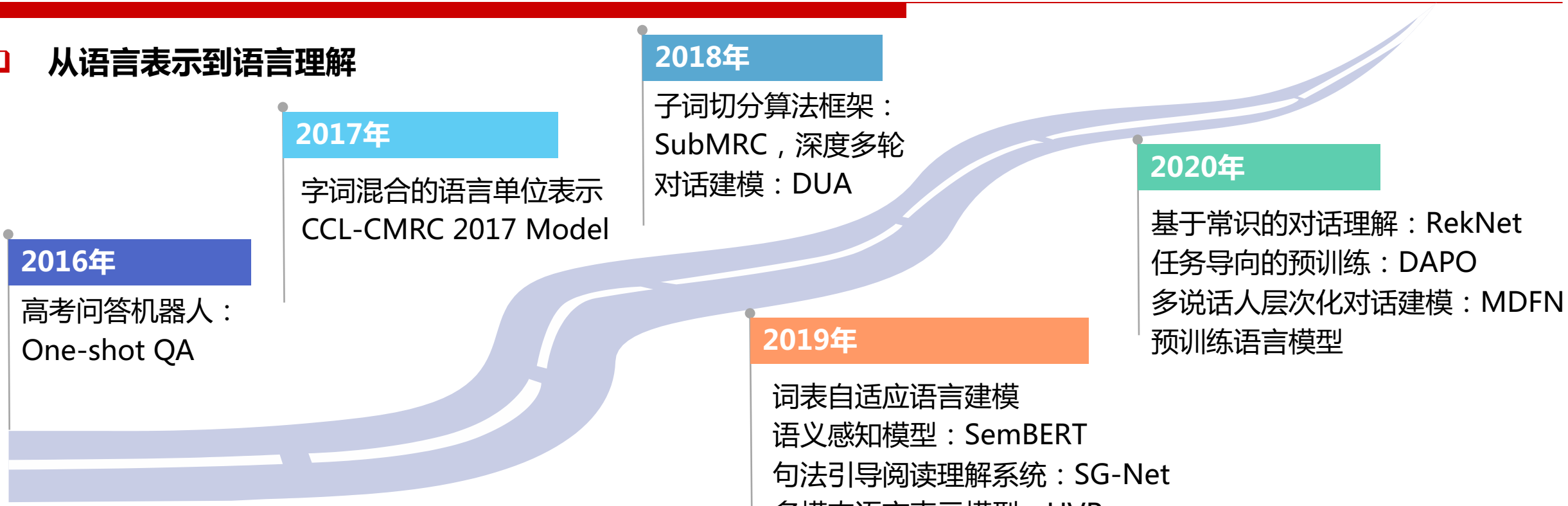
- **2020年 反思：下一阶段的阅读理解？**
  - 常识问题
  - 数值推理
  - 多轮对话理解与回复生成
  - 任务导向的预训练
  - ……

- 作为语言理解核心任务，阅读理解正在快速转型
- 依然存在大量挑战：开放问答、逻辑推理、图表理解等
- 阅读理解与语言模型研究密不可分

综述：Zhang, Zhuosheng, Hai Zhao, and Rui Wang. "Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond." arXiv preprint arXiv:2005.06249 (2020).

研究路线

- **从语言表示到语言理解**



代表论文

COLING 2018: One-shot Learning for Question-Answering in Gaokao History Challenge
COLING 2018: Subword-augmented Embedding for Cloze Reading Comprehension
COLING 2018: Modeling Multi-turn Conversation with Deep Utterance Aggregation
ACL 2019: Open Vocabulary Learning for Neural Chinese Pinyin IME
TASLP: Effective Subword Segmentation for Text Comprehension
AAAI 2020: Semantics-aware BERT for Natural Language Understanding
AAAI 2020: Syntax-Guided Machine Reading Comprehension
ICLR 2020: Neural Machine Translation with Universal Visual Representation

评测经历：

- 2017 年 首届全国中文机器阅读理解评测（CMRC2017）单模型**第一名**
- 2019 年 国际自然语言推理 SNLI 排行榜**第一名**
- 2019 国际权威机器阅读理解评测排行榜 SQuAD 2.0**第一名**
  - 首次以单模型超越**人类基准**
- 2019年国际大规模考试类阅读理解RACE**第一名**
- 2020年获得对话推理Mutual排行榜**第一名**

目录

通用系统架构：Two-stage Solving Architecture

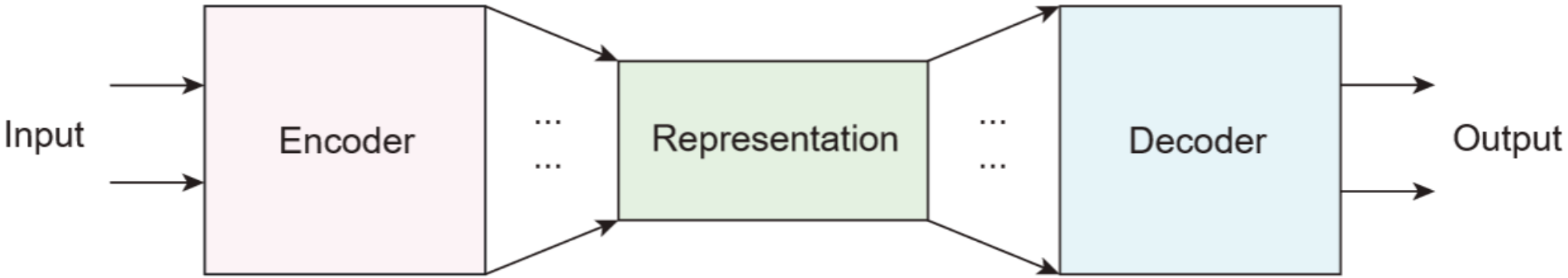
Inspired by Dual process theory of cognition psychology:

the cognitive process of human brains potentially involves two distinct types of procedures:

- **contextualized perception** (reading): gather information in an implicit process
- **analytic cognition** (comprehension): conduct the controlled reasoning and execute goals

Standard MRC system:

- building a CLM as **Encoder**;
- designing ingenious mechanisms as **Decoder** according to task characteristics.



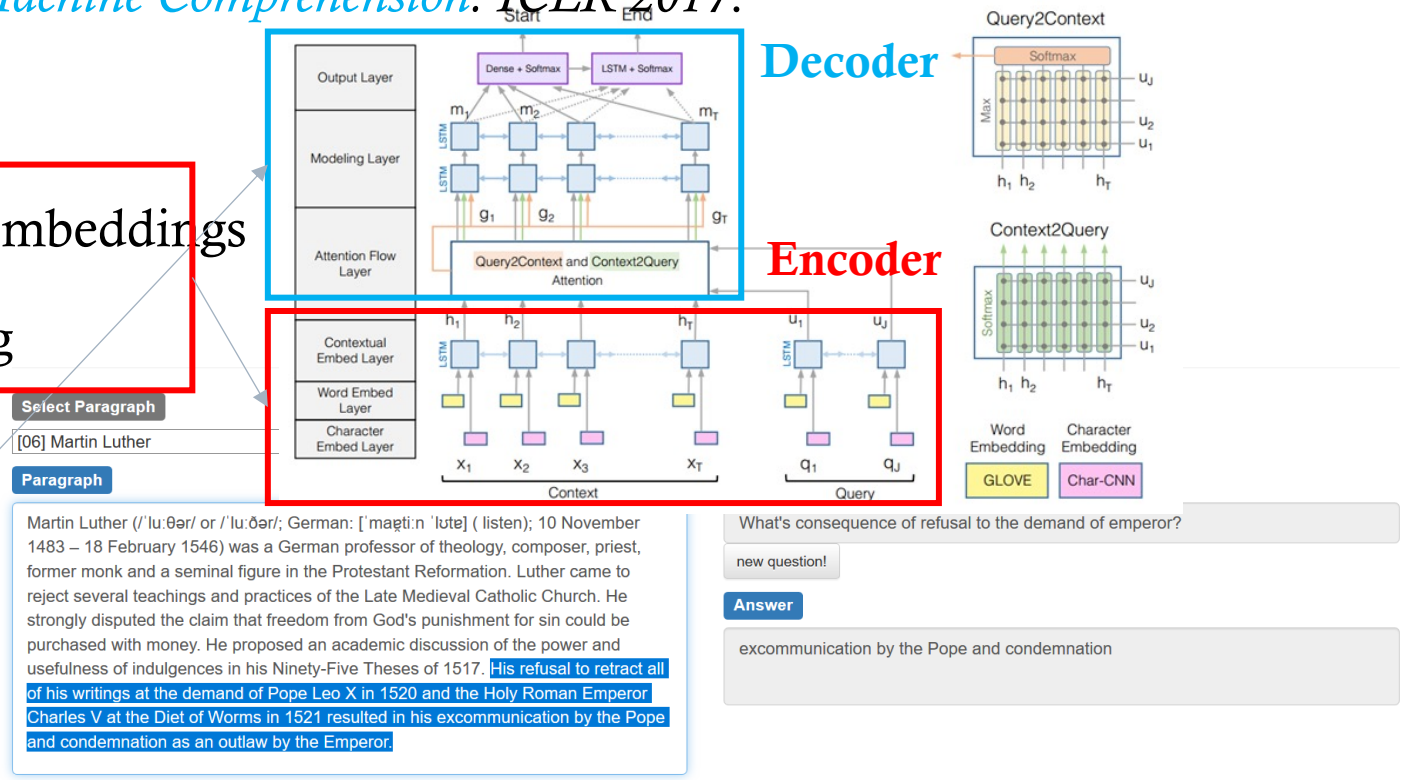
阅读理解经典模型

BiDAF

- *Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. ICLR 2017.*

Hierarchical structure:

- Word + Char level embeddings
- Contextual encoding
- Attention modules
- Answer prediction



Pre-trained models for Fine-tuning

Encoder: Pre-trained Language Models; **Decoder**: most are linear layers.

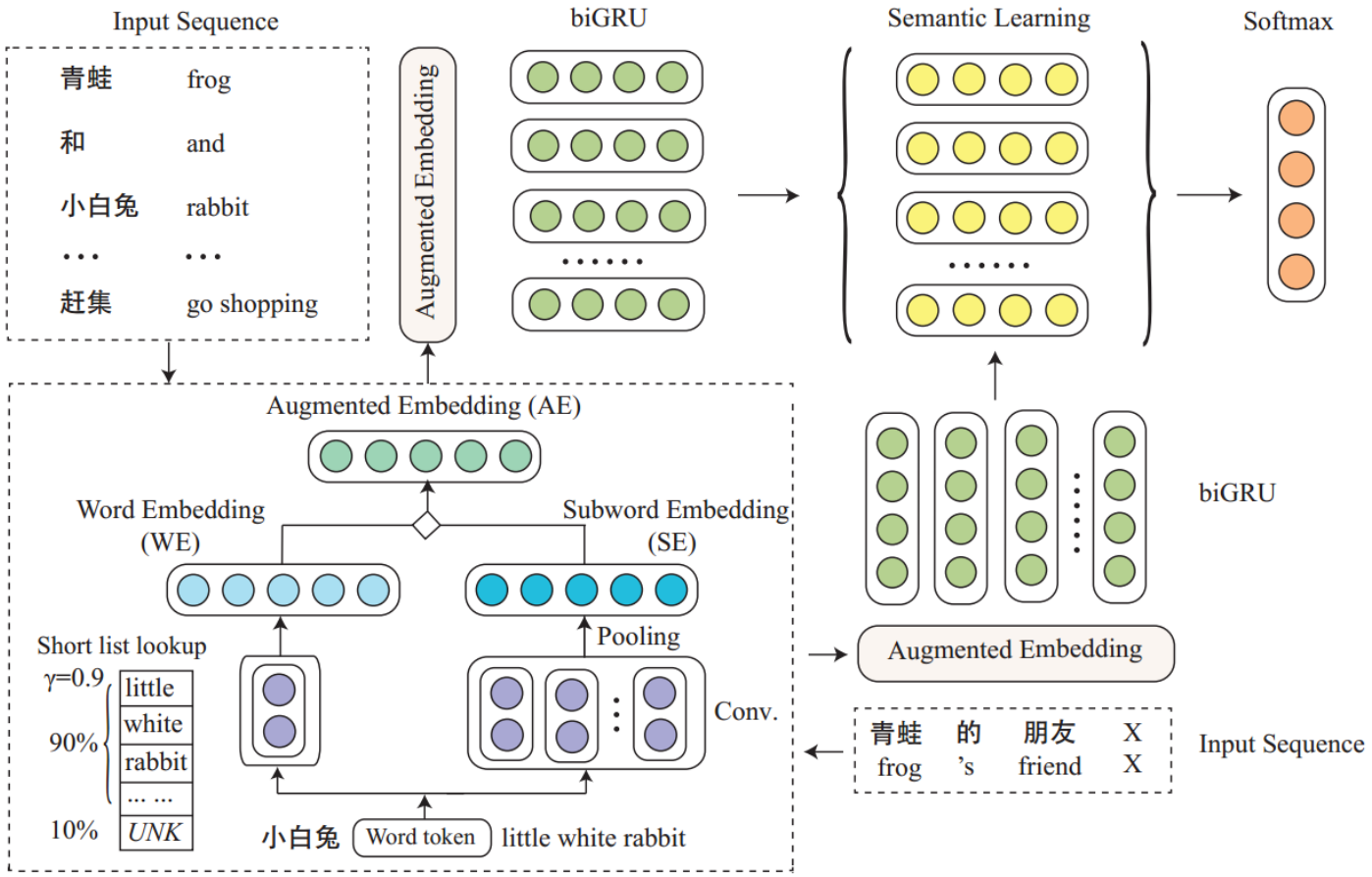
Encoder

- **Multiple Granularity Features**
  - Language Units: word, character, subword.
  - Salient Features: Linguistic features, such as part-of-speech, named entity tags, semantic role labeling tags, syntactic features, and binary Exact Match features.
- **Structured Knowledge Injection (Transformer/GNN)**
  - Linguistic Structures
  - Commonsense
- **Contextualized Sentence Representation**
  - Embedding pretraining

Encoder (our work: language units)

SubMRC: Subword-augmented Embedding

Zhuosheng Zhang, Yafang Huang, Hai Zhao. 2018. Subword-augmented Embedding for Cloze Reading Comprehension. COLING 2018



- Gold answers are often rare words.
- Error analysis shows that early MRC models suffer from out-of-vocabulary (OOV) issues.

We propose:

- Subword-level representation
- Frequency-based short list filtering

We investigate many **subword segmentation algorithms** and propose a unified framework composed of goodness measure and segmentation:

Zhuosheng Zhang, Hai Zhao, Kangwei Ling, Jiangtong Li, Shexia He, Guohong Fu (2019). Effective Subword Segmentation for Text Comprehension. IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP).

Encoder (our work: language units)

SubMRC: Subword-augmented Embedding

Zhuosheng Zhang, Yafang Huang, Hai Zhao. 2018. Subword-augmented Embedding for Cloze Reading Comprehension. COLING 2018

最佳单系统（Best Single System）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	单系统	76.15%	77.73%

最终系统排名

填空类问题（Cloze-style Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	6ESTATES PTE LTD	多系统	81.85%	81.90%
		单系统	75.85%	74.73%
2	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	多系统	78.35%	80.67%
		单系统	76.15%	77.73%
3	南京云思创智信息科技有限公司	多系统	79.20%	80.27%
		单系统	77.15%	77.53%
4	华东师范大学 East China Normal University (ECNU)	多系统	79.45%	79.70%
		单系统	77.95%	77.40%
5	鲁东大学 Ludong University	多系统	77.05%	77.07%
		单系统	74.75%	75.07%
6	武汉大学语言与信息研究中心 Wuhan University (WHU)	单系统	78.20%	76.53%

Best single model in CMRC 2017 shared task

Model	CMRC-2017 Valid	CMRC-2017 Test
Random Guess †	1.65	1.67
Top Frequency †	14.85	14.07
AS Reader †	69.75	71.23
GA Reader	72.90	74.10
SJTU BCMI-NLP †	76.15	77.73
6ESTATES PTE LTD †	75.85	74.73
Xinktech †	77.15	77.53
Ludong University †	74.75	75.07
ECNU †	77.95	77.40
WHU †	78.20	76.53
SAW Reader	**78.95**	**78.80**

Model	PD Valid	PD Test	CFT Test-human
AS Reader	64.1	67.2	33.1
GA Reader	67.2	69.0	36.9
CAS Reader	65.2	68.1	35.0
SAW Reader	**72.8**	**75.1**	**43.8**

Model	CBT-NE Valid	CBT-NE Test	CBT-CN Valid	CBT-CN Test
Human ‡	-	81.6	-	81.6
LSTMs ‡	51.2	41.8	62.6	56.0
MemNets ‡	70.4	66.6	64.2	63.0
AS Reader ‡	73.8	68.6	68.8	63.4
Iterative Attentive Reader ‡	75.2	68.2	72.1	69.2
EpiReader ‡	75.3	69.7	71.5	67.4
AoA Reader ‡	77.8	72.0	72.2	69.4
NSE ‡	78.2	73.2	74.3	71.9
FG Reader ‡	**79.1**	**75.0**	**75.3**	**72.0**
GA Reader ‡	76.8	72.5	73.1	69.6
SAW Reader	78.5	74.9	75.0	71.6

Encoder (our work: salient features)

SemBERT: Semantics-aware BERT

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, Xiang Zhou. 2020. Semantics-aware BERT for Language Understanding. AAAI-2020.

Passage

- *...Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977merger with Radcliffe College.....*

Question

- *What was the name of the leader through the Great Depression and World War II?*

Semantic Role Labeling (SRL)

- [*James Bryant Conant*]$_{\text{ARG0}}$ [*led*]$_{\text{VERB}}$ [*the university*]$_{\text{ARG1}}$ *through* [*the Great Depression and World War II*]$_{\text{ARG2}}$

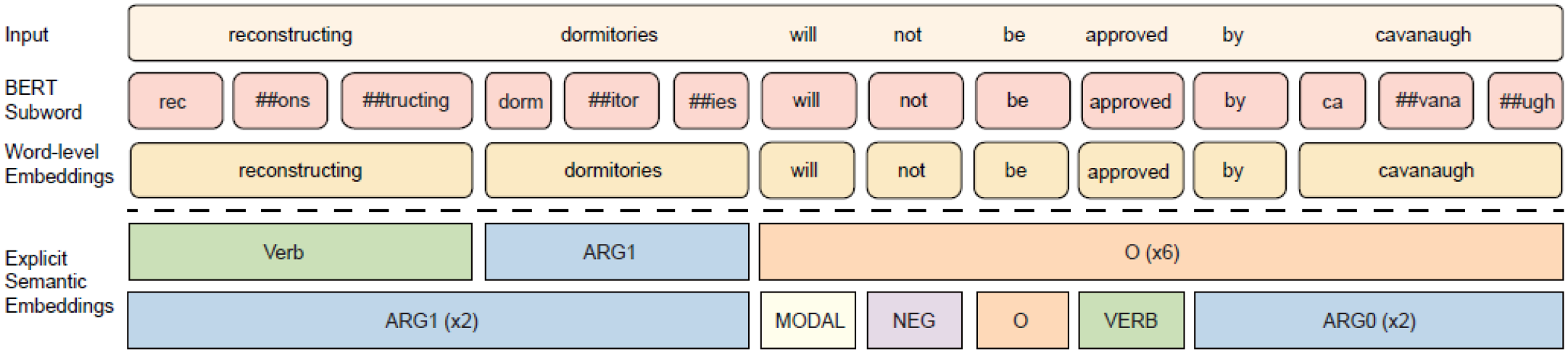
Answer

- *James Bryant Conant*

Encoder (our work: salient features)

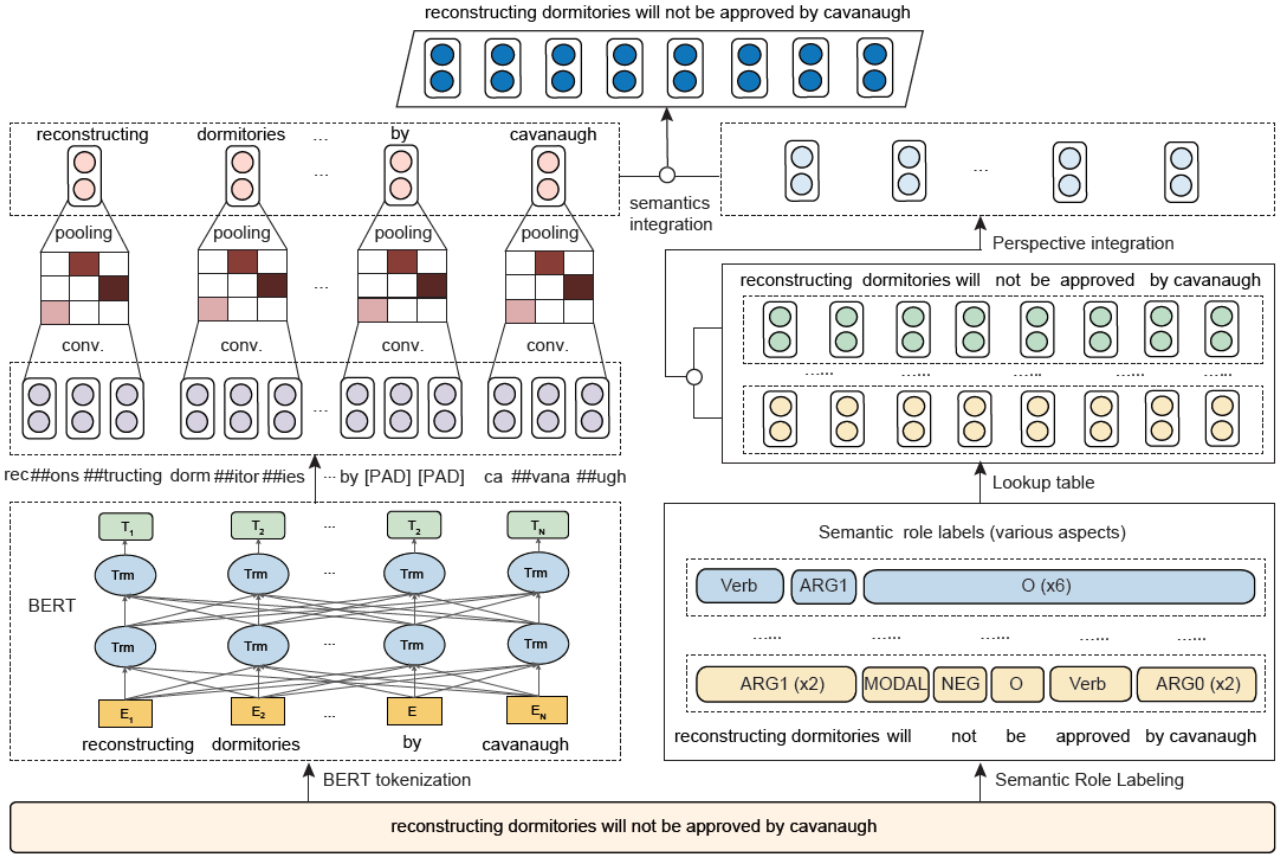
SemBERT: Semantics-aware BERT

- ELMo & BERT: only take Plain contextual features
- SemBERT: introduce Explicit contextual Semantics, Deeper representation?
  - Semantic Role Labeler + BERT encoder



Encoder (our work: salient features)

SemBERT: Semantics-aware



Method	Classification		Natural Language Inference			Semantic Similarity			Score
	CoLA (mc)	SST-2 (acc)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)	- -
Leaderboard (September, 2019)									
ALBERT	69.1	97.1	91.3/91.0	99.2	89.2	93.4	74.2	92.5	89.4
RoBERTa	67.8	96.7	90.8/90.2	98.9	88.2	92.1	90.2	92.2	88.5
XLNET	67.8	96.8	90.2/89.8	98.6	86.3	93.0	90.3	91.6	88.4
In literature (April, 2019)									
BiLSTM+ELMo+Attn	36.0	90.4	76.4/76.1	79.9	56.8	84.9	64.8	75.1	70.5
GPT	45.4	91.3	82.1/81.4	88.1	56.0	82.3	70.3	82.0	72.8
GPT on STILTs	47.2	93.1	80.8/80.6	87.2	69.1	87.7	70.1	85.3	76.9
MT-DNN	61.5	95.6	86.7/86.0	-	75.5	90.0	72.4	88.3	82.2
$BERT_{BASE}$	52.1	93.5	84.6/83.4	-	66.4	88.9	71.2	87.1	78.3
$BERT_{LARGE}$	60.5	94.9	86.7/85.9	92.7	70.1	89.3	72.1	87.6	80.5
Our implementation									
$SemBERT_{BASE}$	57.8	93.5	84.4/84.0	90.9	69.3	88.2	71.8	87.3	80.9
$SemBERT_{LARGE}$	62.3	94.6	87.6/86.3	94.6	84.5	91.2	72.8	87.8	82.9

GLUE 实验结果

Model	EM	F1
#1 BERT + DAE + AoA†	85.9	88.6
#2 SG-Net†	85.2	87.9
#3 BERT + NGM + SST†	85.2	87.7
U-Net (Sun et al. 2018)	69.2	72.6
RMR + ELMo + Verifier (Hu et al. 2018)	71.7	74.2
Our implementation		
$BERT_{LARGE}$	80.5	83.6
$SemBERT_{LARGE}$	82.4	85.2
$SemBERT^{*}_{LARGE}$	84.8	87.9

SQuAD 实验结果

Model	Dev	Test
In literature		
DRCN (Kim et al. 2018)	-	90.1
SJRC (Zhang et al. 2019)	-	91.3
MT-DNN (Liu et al. 2019)†	92.2	91.6
Our implementation		
$BERT_{BASE}$	90.8	90.7
$BERT_{LARGE}$	91.3	91.1
$SemBERT_{BASE}$	91.2	91.0
$SemBERT_{LARGE}$	92.3	91.6

SNLI 实验结果

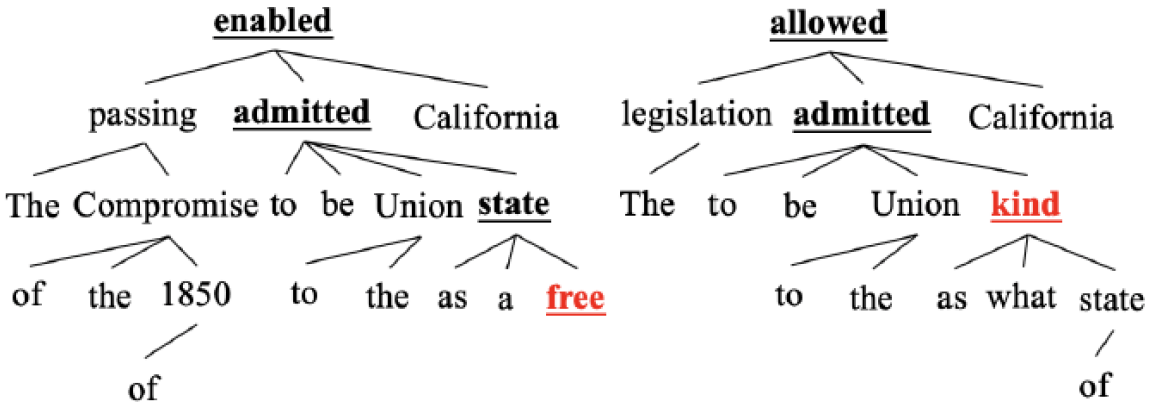
SNLI: The **best** among all submissions.
https://nlp.stanford.edu/projects/snli/
SQuAD2.0: The **best** among all the published work.
GLUE: substantial gains over all the tasks.

Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Transformer

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, Rui Wang*. 2020. SG-Net: Syntax Guided Transformer for Language Representation. TPAMI.*

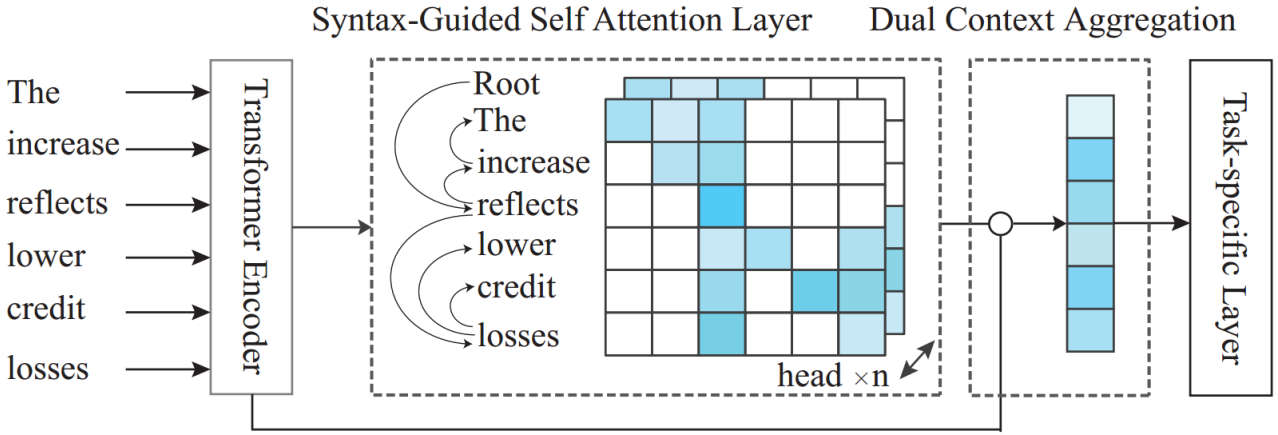
- Passage
  - *The passing of the Compromise of 1850 enabled California to be admitted to the Union as a free state, preventing southern California from becoming its own separate slave state …*
- Question:
  - *The legislation allowed California to be admitted to the Union as what kind of state?*
- Answer:
  - *free*

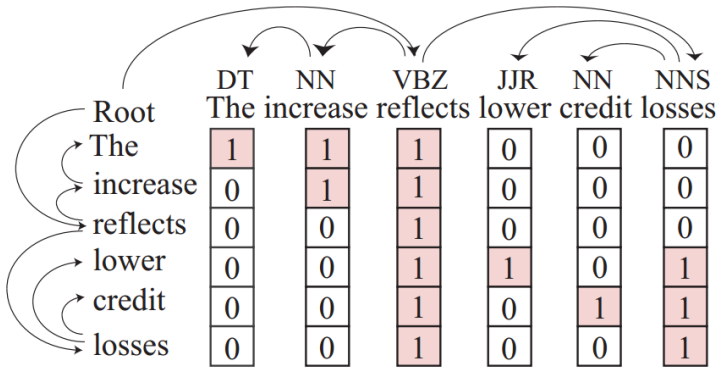


Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Network

- Self-attention network (SAN) empowered Transformer-based encoder
- Syntax-guided self-attention network (SAN)
  - Syntactic dependency of interest (SDOI): regarding each word as a child node
  - SDOI consists all its ancestor nodes and itself in the dependency parsing tree
  - *Pi* : ancestor node set for each i_{th} word; *M* : SDOI mask $\mathcal{M}[i,j] = \begin{cases} 1, & \text{if } j \in P_i \text{ or } j = i \\ 0, & \text{otherwise.} \end{cases}$





Parser: Junru Zhou, Hai Zhao*. 2019.Head-driven Phrase Structure Grammar Parsing on Penn Treebank. ACL 2019, pp.2396–2408.

Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Network

- Our single model (XLNet + SG-Net Verifier) ranks **first**.
- The **first single model** to exceed **human performance**.

Model	Dev EM	Dev F1	Test EM	Test F1
Regular Track				
Joint SAN	69.3	72.2	68.7	71.4
U-Net	70.3	74.0	69.2	72.6
RMR + ELMo + Verifier	72.3	74.8	71.7	74.2
BERT Track				
Human	-	-	86.8	89.5
BERT + DAE + AoA†	-	-	85.9	88.6
BERT + NGM + SST†	-	-	85.2	87.7
BERT + CLSTM + MTL + V†	-	-	84.9	88.2
SemBERT†	-	-	84.8	87.9
Insight-baseline-BERT†	-	-	84.8	87.6
BERT + MMFT + ADA†	-	-	83.0	85.9
$\text{BERT}_{\text{LARGE}}$	-	-	82.1	84.8
Baseline	84.1	86.8	-	-
SG-Net	**85.1**	**87.9**	-	-
+Verifier	**85.6**	**88.3**	**85.2**	**87.9**

Model	RACE-M	RACE-H	RACE
Human Performance			
Turkers	85.1	69.4	73.3
Ceiling	95.4	94.2	94.5
Leaderboard			
DCMN	77.6	70.1	72.3
$\text{BERT}_{\text{LARGE}}$	76.6	70.1	72.0
OCN	76.7	69.6	71.7
Baseline	78.4	70.4	72.6
SG-Net	**78.8**	**72.2**	**74.2**

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 19, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.035	89.897
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language*	86.673	89.147
6 May 21, 2019	XLNet (single model) *Google Brain & CMU*	86.346	89.133
7 May 14, 2019	SG-Net (ensemble) *Shanghai Jiao Tong University*	86.211	88.848
7 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
8	BERT + DAE + AoA (single model)	85.884	88.621

Decoder

- Matching Network:
  - Attention Sum, Gated Attention, Self-matching, Attention over Attention, Co-match Attention, Dual Co-match Attention, etc.
- Answer Pointer:
  - Pointer Network for span prediction
  - Reinforcement learning based self-critical learning to predict more acceptable answers
- Answer Verifier:
  - Threshold-based answerable verification
  - Multitask-style verification
  - External parallel verification
- Answer Type Predictor for multi-type MRC tasks
- Training Objectives

Type	CE	BCE	MSE
Cloze-style	✓		
Span-based	✓		
+ (binary) verification	✓	✓	✓
+ yes/no	✓	✓	✓
+ count	✓		
Multi-choice	✓		

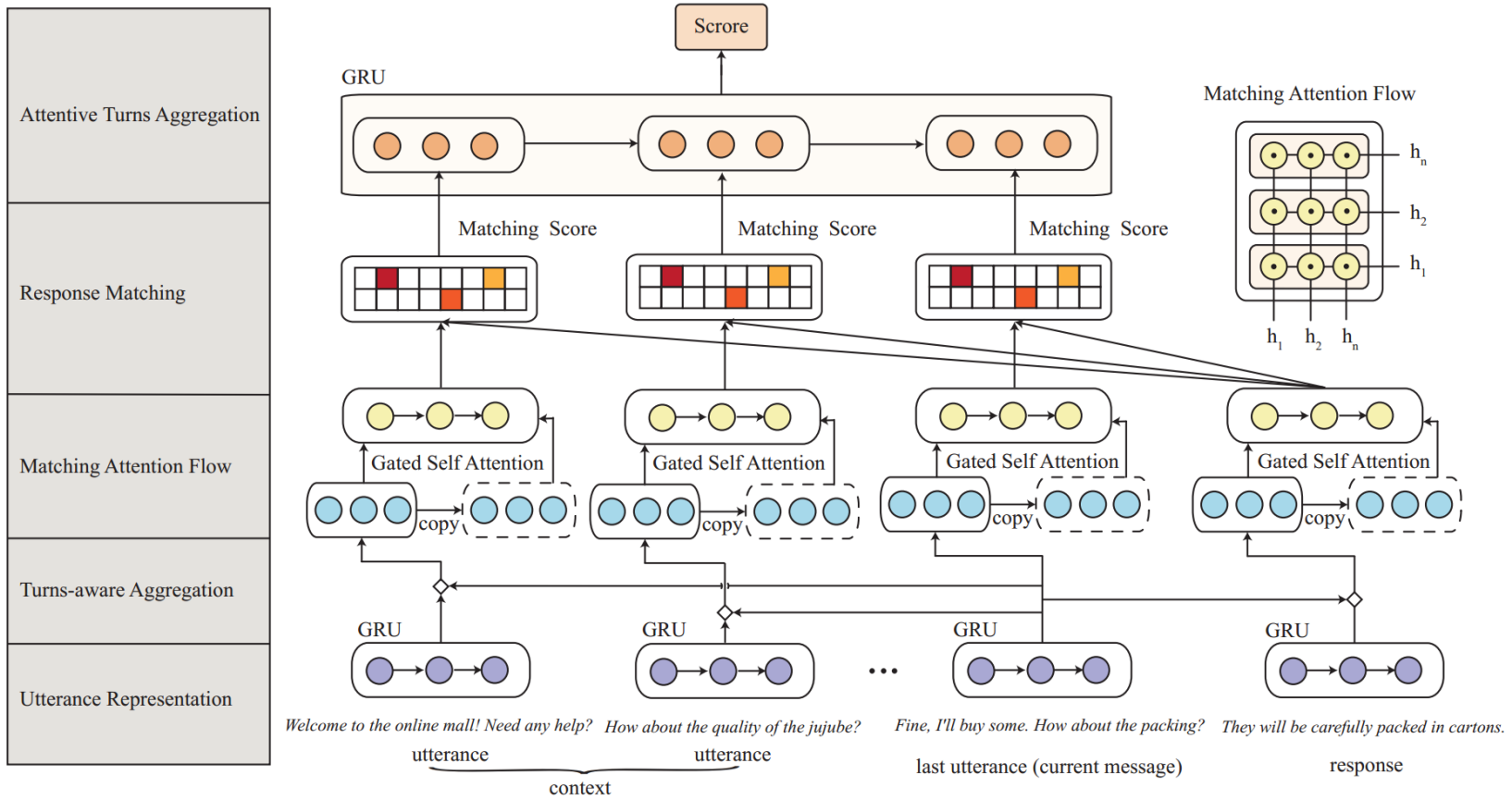
Decoder (our work: Deep Utterance Aggregation)

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. COLING 2018.*

- Challenge: **long utterances, multiple intentions, topic shift**, etc.
- Aim: recognize the **key information** from complex dialogue history
- Solution: deep utterance aggregation framework (**DUA**)
- Corpus: a new **E-commerce Dialogue Corpus**

Decoder (our work: Deep Utterance Aggregation)

- Capture the main information in each utterance (**self attention,** first introduced)
- Model the **information flow through the utterances** in dialogue history
- Match the relationship **between utterance and candidate response**



Highlight the importance of **the last utterance**.

Decoder (our work: Deep Utterance Aggregation)

Appeared in the Google Scholar 2021 h5-index list, **top 1.2%,** 16/1282 in COLING in the last 5 years.

（近五年COLING高引论文前16）

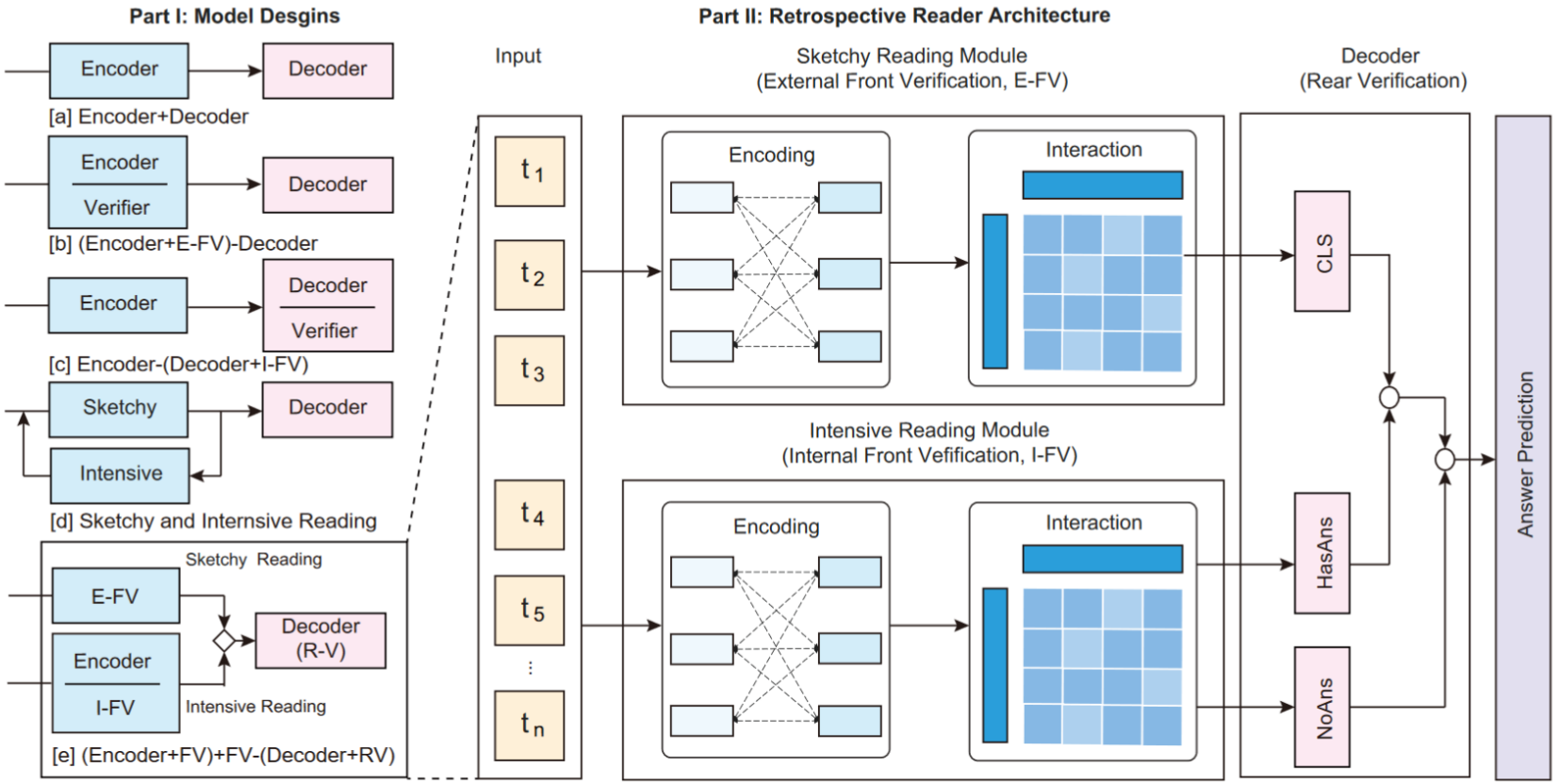
Title / Author	Cited by	Year
Contextual String Embeddings for Sequence Labeling A Akbik, D Blythe, R Vollgraf Proceedings of the 27th International Conference on Computational …	802	2018
Effective LSTMs for Target-Dependent Sentiment Classification. D Tang, B Qin, X Feng, T Liu COLING, 3298-3307	522	2016
Automatic Detection of Fake News V Pérez-Rosas, B Kleinberg, A Lefevre, R Mihalcea Proceedings of the 27th International Conference on Computational …	376	2018
Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. P Zhou, Z Qi, S Zheng, J Xu, H Bao, B Xu COLING, 3485-3495	365	2016
A Survey on Recent Advances in Named Entity Recognition from Deep Learning models V Yadav, S Bethard Proceedings of the 27th International Conference on Computational …	301	2018
SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. E Cambria, S Poria, R Bajpai, BW Schuller COLING, 2666-2677	287	2016
A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. S Poria, E Cambria, D Hazarika, P Vij COLING, 1601-1612	245	2016
Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. X Wang, W Jiang, Z Luo COLING, 2428-2437	241	2016
Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. L Mou, Y Song, R Yan, G Li, L Zhang, Z Jin COLING, 3349-3358	187	2016
Sentence Similarity Learning by Lexical Decomposition and Composition. Z Wang, H Mi, A Ittycheriah COLING, 1340-1349	183	2016
Recurrent Dropout without Memory Loss. S Semeniuta, A Severyn, E Barth COLING, 1757-1766	178	2016
Neural Paraphrase Generation with Stacked Residual LSTM Networks. A Prakash, SA Hasan, K Lee, VV Datla, A Qadir, J Liu, O Farri COLING, 2923-2934	176	2016
Improved relation classification by deep recurrent neural networks with data augmentation. Y Xu, R Jia, L Mou, G Li, Y Chen, Y Lu, Z Jin COLING, 1461-1470	173	2016
Diachronic word embeddings and semantic shifts: a survey A Kutuzov, L Øvrelid, T Szymanski, E Velldal Proceedings of the 27th International Conference on Computational …	163	2018
Attending to Characters in Neural Sequence Labeling Models. M Rei, GKO Crichton, S Pyysalo COLING, 309-318	153	2016
Modeling Multi-turn Conversation with Deep Utterance Aggregation Z Zhang, J Li, P Zhu, H Zhao, G Liu Proceedings of the 27th International Conference on Computational …	143	2018
A Survey of Domain Adaptation for Neural Machine Translation C Chu, R Wang Proceedings of the 27th International Conference on Computational …	134	2018
Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification. A Dahou, S Xiong, J Zhou, MH Haddoud, P Duan COLING, 2418-2427	131	2016
Automated Fact Checking: Task Formulations, Methods and Future Directions J Thorne, A Vlachos Proceedings of the 27th International Conference on Computational …	130	2018
Semi-supervised Word Sense Disambiguation with Neural Models. D Yuan, J Richardson, R Doherty, C Evans, E Altendorf COLING, 1374-1385	130	2016

https://scholar.google.com/citations?hl=en&vq=eng_computationallinguistics&view_op=list_hcore&venue=6AfzgED5a7MJ.2021

Decoder (our work: answer verifier)

- Retro-Reader

Zhuosheng Zhang, Junjie Yang, Hai Zhao (2021). Retrospective Reader for Machine Reading Comprehension. AAAI 2021.



- Multitask Internal Verification
- Parallel External Verification
- Rear Verification

Decoder (our work: answer verifier)

- Retro-Reader

SOTA results on SQuAD 2.0 and NewsQA

Passage:
Southern California consists of a heavily developed urban environment, home to some of the largest urban areas in the state, along with vast areas that have been left undeveloped. It is the third most populated megalopolis in the United States, after the Great Lakes Megalopolis and the Northeastern megalopolis. Much of southern California is famous for its large, spread-out, suburban communities and use of automobiles and highways...

Question:
What are the second and third most populated megalopolis after Southern California?

Answer:
Gold: ⟨no answer⟩
ALBERT (+TAV): Great Lakes Megalopolis and the Northeastern megalopolis.
Retro-Reader over ALBERT: ⟨no answer⟩
$score_{has} = 0.03, score_{na} = 1.73, \lambda = -0.98$

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
5 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019

目录

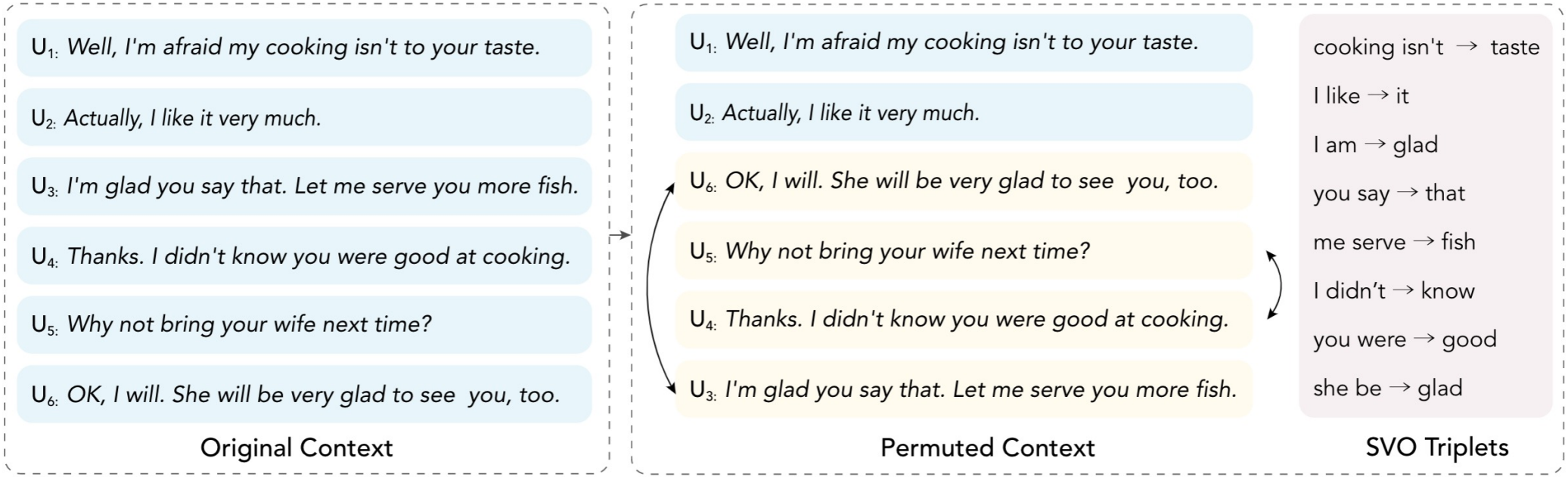
结构化对话预训练

Zhuosheng Zhang, Hai Zhao, 2021. Structural Pre-training for Dialogue Comprehension. ACL 2021.*

- **Background:** How to train language models on dialogue scenarios
  - open-domain pre-training
  - domain-adaptive post-training
- **Motivation:**
  - The pre-trained models handle the whole input text as a linear sequence of successive tokens
  - It is challenging to effectively capture task-related knowledge from dialogue texts
  - Dialogue contexts are composed of many utterances from different speakers
  - Dialogues are rich in complex discourse structures and correlations

结构化对话预训练

- **SPIDER**: Structural Pre-trained Dialogue Reader
  - **utterance order restoration:** predicts the order of the permuted utterances
  - **sentence backbone regularization:** improve the factual correctness of SVO triples
- Efficiently and explicitly model the coherence among utterances and the key facts in utterances



结构化对话预训练

Model	Ubuntu Corpus			Douban Conversation Corpus						E-commerce Corpus		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
SMN	72.6	84.7	96.1	52.9	56.9	39.7	23.3	39.6	72.4	45.3	65.4	88.6
DUA	75.2	86.8	96.2	55.1	59.9	42.1	24.3	42.1	78.0	50.1	70.0	92.1
DAM	76.7	87.4	96.9	55.0	60.1	42.7	25.4	41.0	75.7	-	-	-
IoI	79.6	89.4	97.4	57.3	62.1	44.4	26.9	45.1	78.6	-	-	-
MSN	80.0	89.9	97.8	58.7	63.2	47.0	29.5	45.2	78.8	60.6	77.0	93.7
MRFN	78.6	88.6	97.6	57.1	61.7	44.8	27.6	43.5	78.3	-	-	-
SA-BERT	85.5	92.8	98.3	**61.9**	**65.9**	**49.6**	**31.3**	48.1	**84.7**	70.4	**87.9**	98.5
Multi-task Fine-tuning												
BERT	81.7	90.4	97.7	58.8	63.1	45.3	27.7	46.4	81.8	61.7	81.1	97.0
+ SPIDER	83.1	91.3	98.0	59.8	63.8	45.9	28.5	48.7	82.6	62.6	82.7	97.1
Domain Adaptive Post-training												
BERT	85.7	93.0	98.5	60.5	64.7	47.4	29.1	47.8	84.9	66.4	84.8	97.6
+ SPIDER	**86.9**	**93.8**	**98.7**	60.9	65.0	47.5	29.6	**48.8**	83.6	**70.8**	85.3	**98.6**

基于解耦的图建模

Siru Ouyang#, Zhuosheng Zhang#, Hai Zhao, 2021. Dialogue Graph Modeling for Conversational Machine Reading. Findings of ACL 2021.*

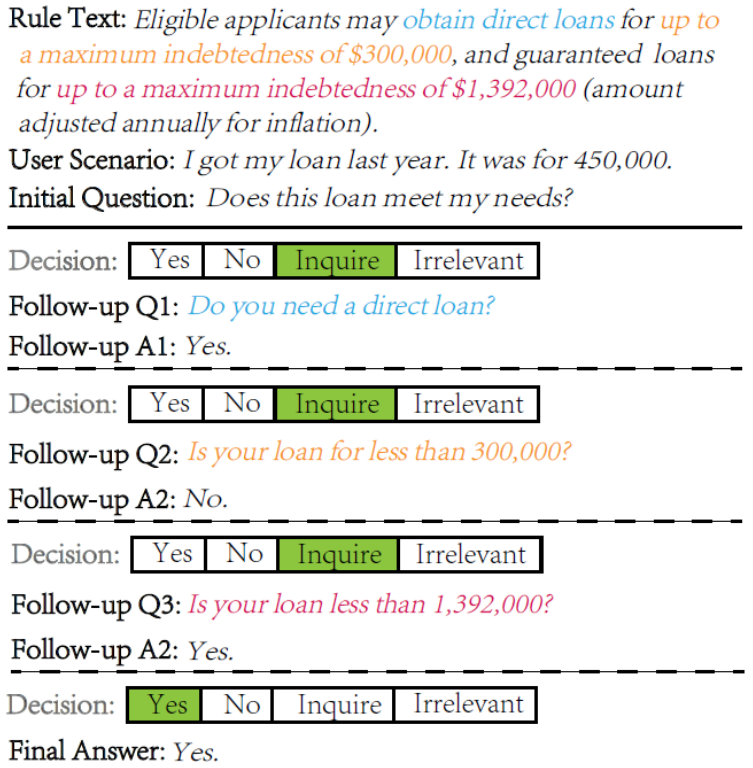
- Task: Conversational Machine Reading

  Input x = (r, s, q, h)

  - r: Rule Text
  - s: User Scenario
  - q: Initial Question
  - h: Dialogue History

  Output (divided into two subtasks):

  - A decision ∈ (yes, no, inquire, irrelevant)
  - If *inquire*, ask a follow-up question



An example taken from the ShARC (Saeidi et al., 2018) benchmark

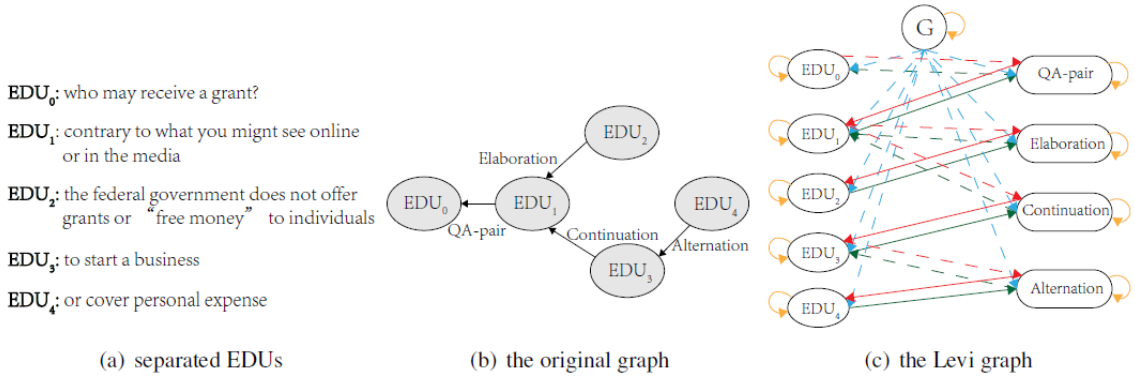
基于解耦的图建模

- Interpreting rule document
  - Identify rule conditions
  - Discourse relations among rule conditions
  - Interactions among all the elements (scenario, question, etc.)
- Make decisions as the conversation flows
  - Track fulfillment over identified rule conditions
  - jointly consider fulfillment states to make the final decision

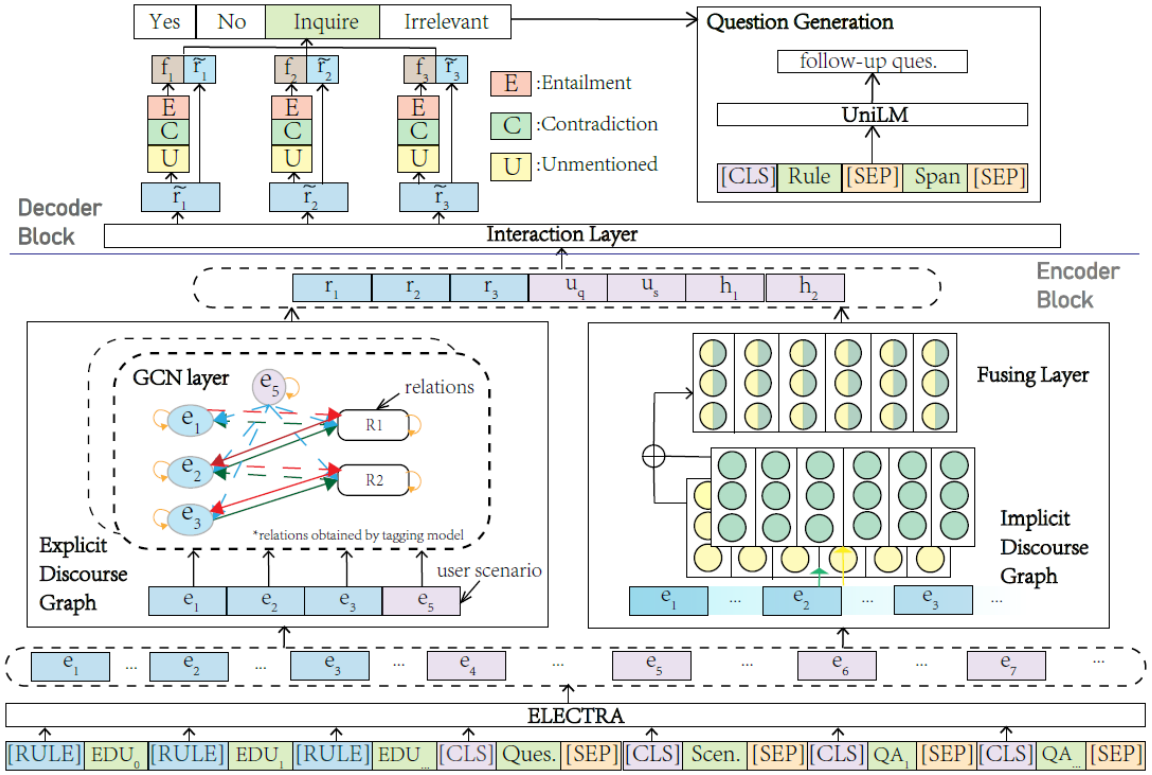
基于解耦的图建模

- Explicit Discourse Graph:

  injects the discourse relations via open-source tagging tool



Using RGCN models to encode the graph.



基于解耦的图建模

Model	Dev Set				Test Set			
	Decision Making		Question Gen.		Decision Making		Question Gen.	
	Micro	Macro	BLEU1	BLEU4	Micro	Macro	BLEU1	BLEU4
NMT (Saeidi et al., 2018)	-	-	-	-	44.8	42.8	34.0	7.8
CM (Saeidi et al., 2018)	-	-	-	-	61.9	68.9	54.4	34.4
BERTQA (Zhong and Zettlemoyer, 2019)	68.6	73.7	47.4	54.0	63.6	70.8	46.2	36.3
UcraNet (Verma et al., 2020)	-	-	-	-	65.1	71.2	60.5	46.1
BiSon (Lawrence et al., 2019)	66.0	70.8	46.6	54.1	66.9	71.6	58.8	44.3
E^3 (Zhong and Zettlemoyer, 2019)	68.0	73.4	67.1	53.7	67.7	73.3	54.1	38.7
EMT (Gao et al., 2020a)	73.2	78.3	67.5	53.2	69.1	74.6	63.9	**49.5**
DISCERN (Gao et al., 2020b)	74.9	79.8	65.7	52.4	73.2	78.3	**64.0**	49.1
DGM (ours)	**78.6**	**82.2**	**71.8**	**60.2**	**77.4**	**81.2**	63.3	48.4

Evaluation Metrics

- Decision Making: Micro-accuracy and Macro-accuracy
- Question Generation: BLEU1 and BLEU4

事实驱动知识推理

Siru Ouyang#, Zhuosheng Zhang#, Hai Zhao*, 2021. Fact-driven Logical Reasoning.

- Task: Logical Reasoning
  - Challenges: entity-aware commonsense, perception of facts or events.
  - Logical supervision is rarely available during language model pre-training.

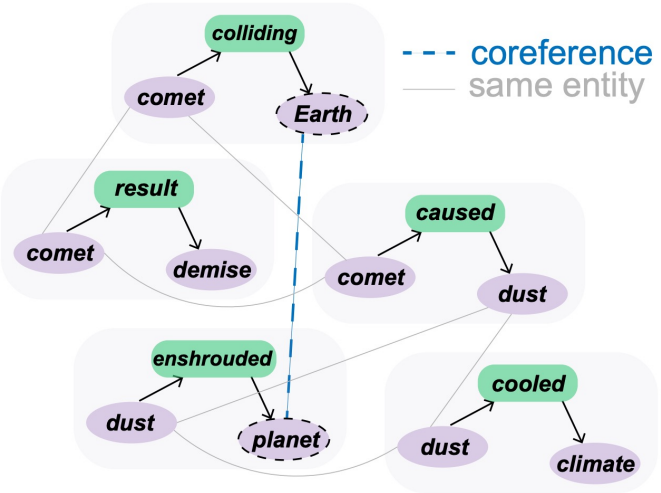
Question	*Passage*	*Answer*
Example 1 From this we know	Xiao Wang is taller than Xiao Li, Xiao Zhao is taller than Xiao Qian, Xiao Li is shorter than Xiao Sun, and Xiao Sun is shorter than Xiao Qian.	√ A. Xiao Li is shorter than Xiao Zhao. B. Xiao Wang is taller than Xiao Zhao. C. Xiao Sun is shorter than Xiao Wang. D. Xiao Sun is taller than Xiao Zhao.
Example 2 Which one of the following statements, most seriously weakens the argument?	.... A large enough comet colliding with Earth could have caused a cloud of dust that enshrouded the planet and cooled the climate long enough to result in the dinosaurs' demise.	A. Many other animal species from same era did not become extinct at the same time the dinosaurs did. B. It cannot be determined from dinosaur skeletons whether the animals died from the effects of a dust cloud. C. The consequences for vegetation and animals of a comet colliding with Earth are not fully understood. √ D. Various species of animals from the same era and similar to them in habitat and physiology did not become extinct.

事实驱动知识推理

- Natural logic units would be the group of backbone constituents of the sentence such as subject, verb and object that cover both global and local knowledge pieces.

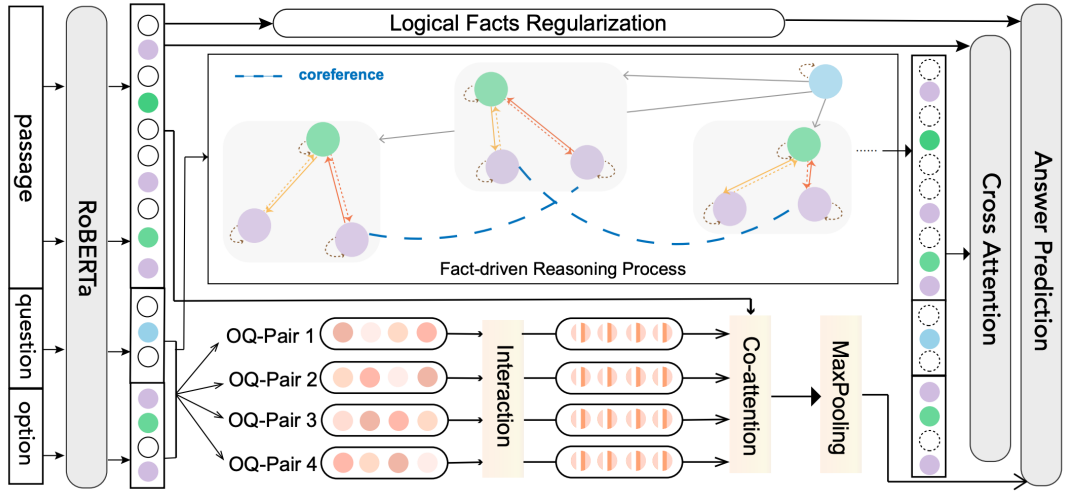
Definition 1 ***(Fact Unit)*** *Given an triplet $T = \{E_1, R, E_2\}$, where E_1 and E_2 are entities, P is the predicate between them, a fact unit F is the set of all entities in T and their corresponding relations.*

Definition 2 ***(Supergraph)*** *A supergraph is a structure made of fact units (regarded as subgraphs) as the vertices, and the coreference relations as undirected edges.*



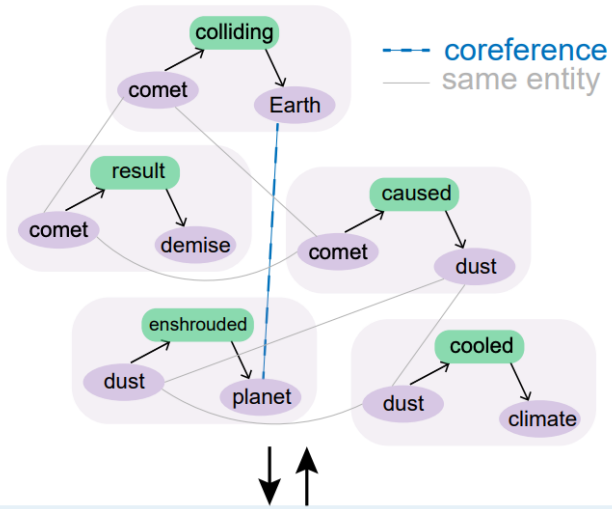
事实驱动知识推理

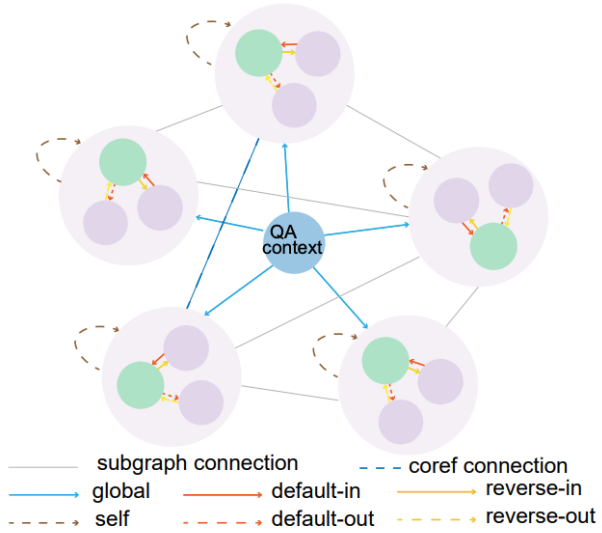
- Supergraph Modeling
  - Build a supergraph on our newly defined fact units
  - Question-Option-aware Interaction
  - Logical Fact Regularization



A large enough comet colliding with Earth could have caused a cloud of dust that enshrouded the planet and cooled the climate long enough to result in the dinosaurs' demise.

comet colliding → Earth
comet caused → dust
dust enshrouded → planet
dust cooled → climate
comet result → demise





Which one of the following ....., most seriously weakens the argument?
Various species of animals from the same era as dinosaurs and similar to them ... did not become extinct when the dinosaurs did.

事实驱动知识推理

- Dramatic improvements on the logical reasoning benchmarks
- FOCAL REASONER makes better use of logical structure inherent in the given context to perform reasoning than existing methods.

Model	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
Human [3]	-	63.00	57.10	67.20	-	86.00
BERT-Large [3]	53.80	49.80	72.00	32.30	34.10	31.03
XLNet-Large [3]	62.00	56.00	75.70	40.50	-	-
RoBERTa-Large [3]	62.60	55.60	75.50	40.00	35.02	35.33
DAGN [6]	65.20	58.20	76.14	44.11	35.48	38.71
DAGN (Aug) [6]	65.80	58.30	75.91	44.46	36.87	39.32
FOCAL REASONER	**66.80**	**58.90**	**77.05**	**44.64**	**41.01**	**40.25**

Model	MuTual						MuTual^{plus}					
	Dev Set			Test Set			Dev Set			Test Set		
	R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR	R_4@1	R_4@2	MRR
RoBERTa$_{base}$ [38]	69.5	87.8	82.4	71.3	89.2	83.6	62.2	85.3	78.2	62.6	**86.6**	78.7
-MC [38]	69.3	88.7	82.5	68.6	88.7	82.2	62.1	83.0	77.8	64.3	84.5	79.2
FOCAL REASONER	**73.4**	**90.3**	**84.9**	**72.7**	**91.0**	**84.6**	**63.7**	**86.1**	**79.1**	**65.5**	84.3	**79.7**

事实驱动知识推理

- An example of how our model reasons to get the final answer

A recent survey in a key middle school showed that high school students in this school have a special preference for playing football, and it far surpasses other balls.The survey also found that students who regularly play football are better at academic performance than students who do not often play football.This shows that often playing football can improve students' academic performance.

✓ A. Only high school students who are ranked in the top 30% of grades can often play football.
B. Regular football can exercise and maintain a strong learning energy.
C. Often playing football delays the study time.
D. Research has not proved that playing football can contribute to intellectual development.

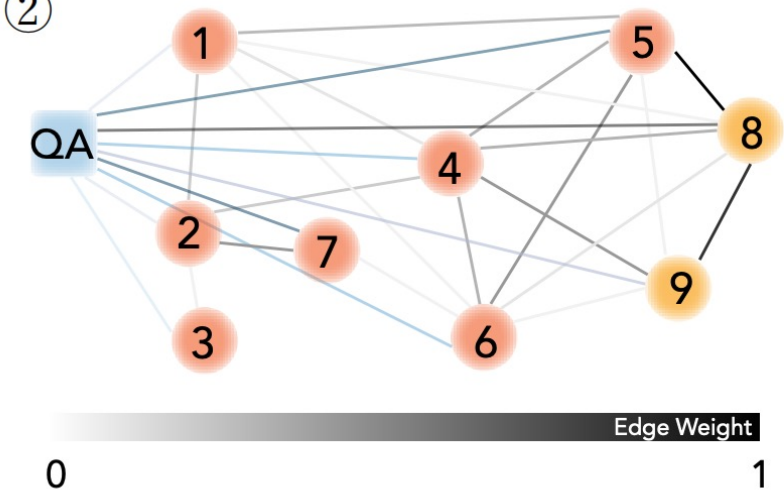
Which of the following can weaken the above conclusion most?

①

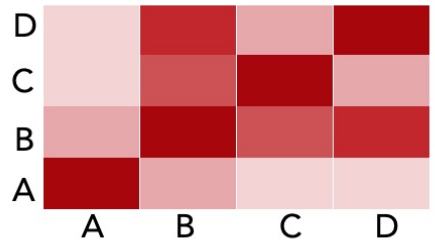
1. students have preferences
2. preference playing football
3. it surpasses balls
4. who play football
5. students better performance
6. who !play football
7. playing improve performance
8. students rank 30%
9. students play football

Fact Units

②



③



Option Similarity Matrix after Interaction

✓ A: 4.1918
B: -5.3050
C: -12.3718
D: -6.9722

Sources

Our Survey Papers:

[1] Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond

Paper Link: https://arxiv.org/abs/2005.06249

[2] Advances in Multi-turn Dialogue Comprehension: A Survey

Paper Link: https://arxiv.org/abs/2103.03125

Our codes are publicly available at: https://github.com/cooelf

Q & A

Thank You !