



我的机器阅读理解之路 — 科研历练中的自我认知与坚持

张倬胜

上海交通大学

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：

- 机器阅读理解榜单与评测**第一名**：
SQuAD2.0、RACE、MuTual、CMRC、SNLI等
- 上海交通大学研究生学术之星
- 中国大学生年度人物提名奖
- 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	GESTATES 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

https://nealclly.github.io/MuTual-leaderboard/

MuTual Challenge

What is MuTual?

MuTual is a retrieval-based dataset for **Multi-Turn** dialogue reasoning, which is modified from Chinese high school English listening comprehension test data. The goal of the MuTual challenge is to evaluate the reasoning ability in chatbots.

[MuTual Paper](#)

Download

[Download Training Set](#)

[Download Dev Set](#)

[Download Test Set](#)

Evaluation

Once you are satisfied with your model performance on the dev set, you are encouraged to send your decode output to cuileyang@zju.edu.cn with your dev performance and methods to get the official scores on the test sets.

Leaderboard

Rank	Model	R@1	R@2	MRR	R@1	R@2	MRR
	Human Performance (non-native speaker)	0.938	0.971	0.964	0.930	0.972	0.961
1	MDFN SJTU & Huawei Noah's Ark Lab	0.916	0.984	0.956	-	-	-
2	GRN-v2 Anonymous	0.915	0.983	0.954	0.841	0.957	0.913
3	GRN-v1 Anonymous	0.903	0.976	0.947	-	-	-
4	UMN Anonymous	0.870	0.973	0.930	-	-	-
5	RoBERTa + OCN Pattern Recognition Center, WeChat AI	0.867	0.958	0.926	-	-	-
6	RoBERTa+ Northeastern University	0.825	0.953	0.904	-	-	-
7	DRRC-1 PKU	0.771	0.914	0.869	-	-	-
8	RoBERTa ZJU & MSRA & Westlake	0.713	0.892	0.836	0.626	0.866	0.787

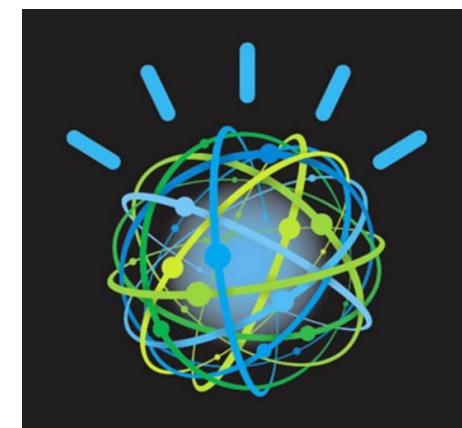
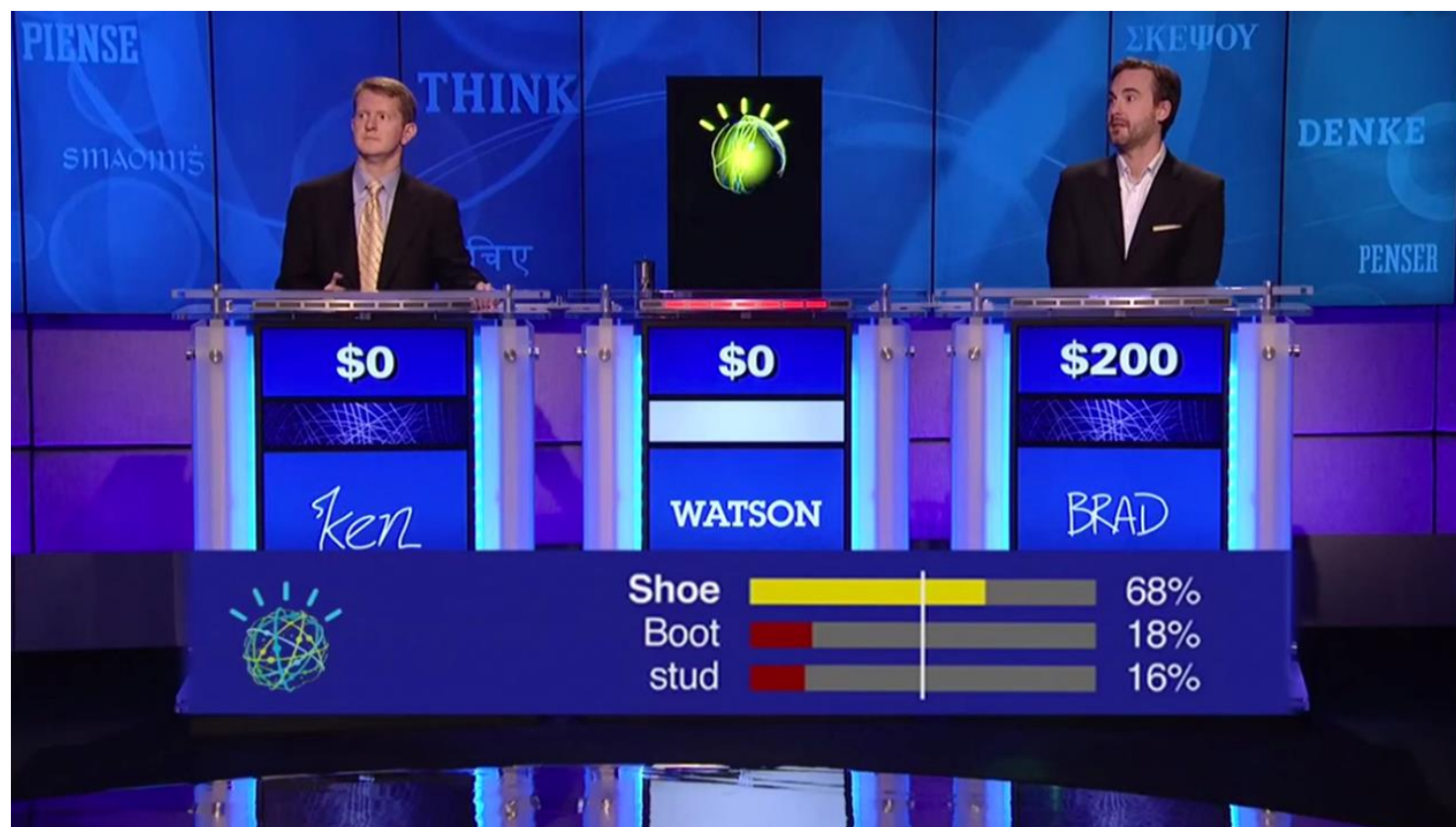
目录

- ❑ 研究历程
- ❑ 研究感悟

为什么做阅读理解?

- ❑ 让机器能听会说，能理解会思考，是一件很cool的事情!

IBM Watson Beats Human Champions!



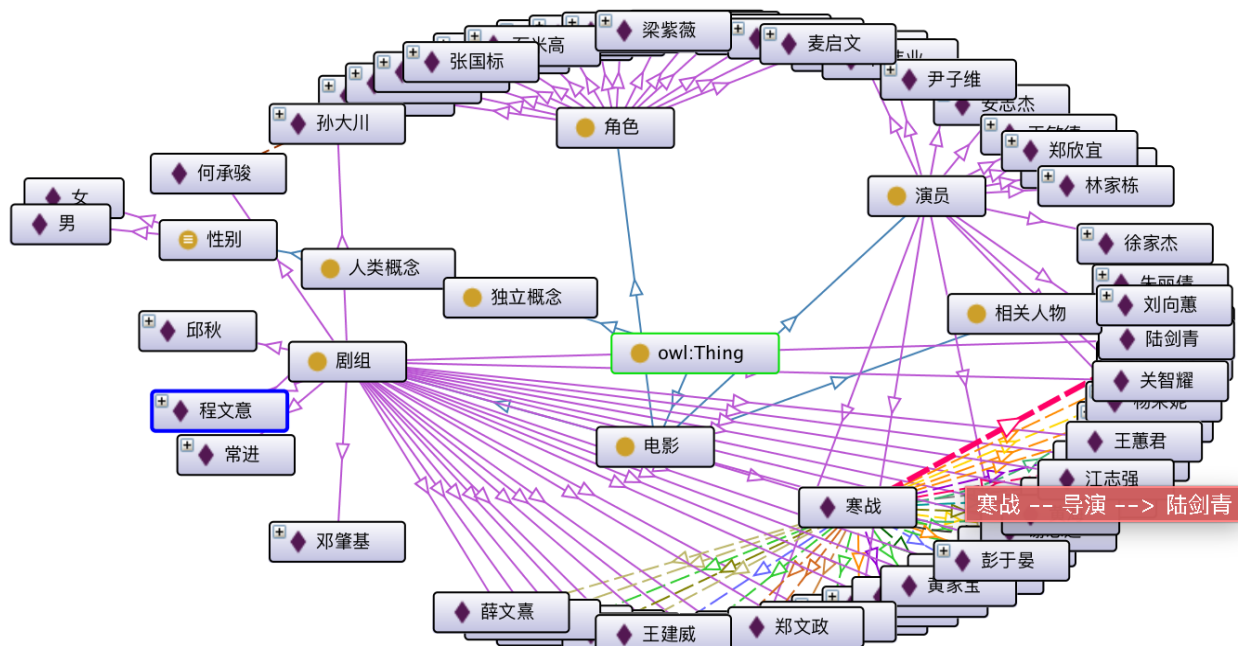
2011 Jeopardy



为什么做阅读理解?

□ 让机器能听会说, 能理解会思考, 是一件很cool的事情!

● 尝试搭建了一套基于知识图谱的中文问答系统



为什么做阅读理解？

□ 让机器能听会说，能理解会思考，是一件很cool的事情！

● 尝试搭建了一套基于知识图谱的中文问答系统

问题:寒战的导演是谁

答案:寒战的导演是陆剑青、梁乐民

问题:寒战的主演是哪些人

答案:寒战的主演是梁家辉、郭富城、杨采妮、李治廷、彭于晏

问题:郭富城的女友是谁

答案:郭富城的女友是方媛

问题:李家俊的父亲在剧中担任什么职位

答案:李家俊的父亲李文斌担任警务处现任行动副处长，为警队鹰派人物，警队内部及前线员工，均视他为下任处长之选。李立即指挥营救，并将此次行动代号为“寒战”。适逢警队一哥外访未返，总部会议室内，李文彬成为香港治安最高决策人。

问题:李家俊的配音是谁

答案:李家俊的配音是何承骏

问题:梁家辉毕业于哪所大学

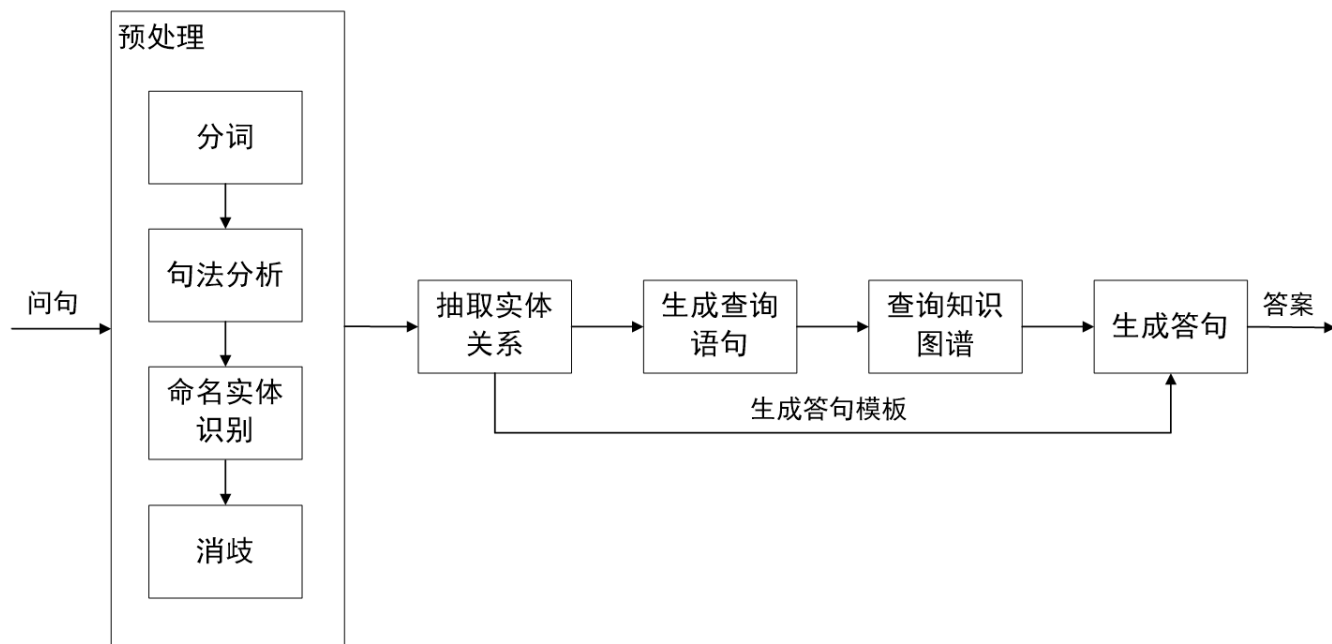
答案:梁家辉毕业香港理工大学

为什么做阅读理解?

□ 让机器能听会说, 能理解会思考, 是一件很cool的事情!

- 尝试搭建了一套基于知识图谱的中文问答系统
- 遇到的问题: 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

语言理解
路还很长!



问题:刘德华出生于哪里

消歧:刘德华诞生哪里

问题:刘德华的老婆是谁

消歧:刘德华的妻子是谁

问题:刘德华的哥们有哪些

消歧:刘德华的兄弟有哪些

为什么做阅读理解？

- ❑ **让机器能听会说，能理解会思考，是一件很cool的事情！**
 - 尝试搭建了一套**基于知识图谱的中文问答系统**
 - 遇到的问题：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
- ❑ 阅读理解任务从2015年开始逐渐得到广泛关注（类似中高考阅读理解题型）
 - 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
- ❑ **第一个吃螃蟹：实验室**第一个**从事机器阅读理解研究**
 - 从“实验翻车、论文难产”探索门路

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

研究路线

2016年

高考问答机器人：
One-shot QA

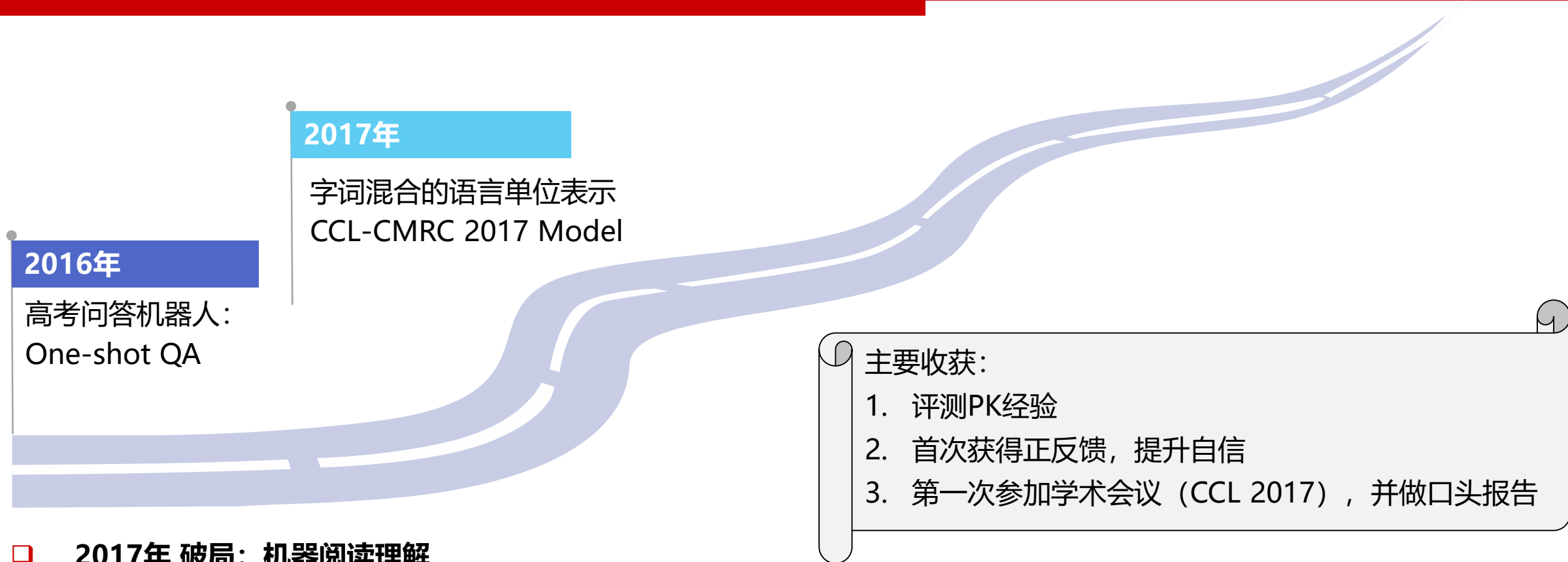
主要收获：

1. 相对完整的科研训练
2. 问题解决能力的提升
3. 对“挫败”的体会与认识

□ 2016年初探：高考机器人（历史）

- 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
- 论文难产：从2017年初投到2018年中旬
- 总结：工作超前，应脚踏实地

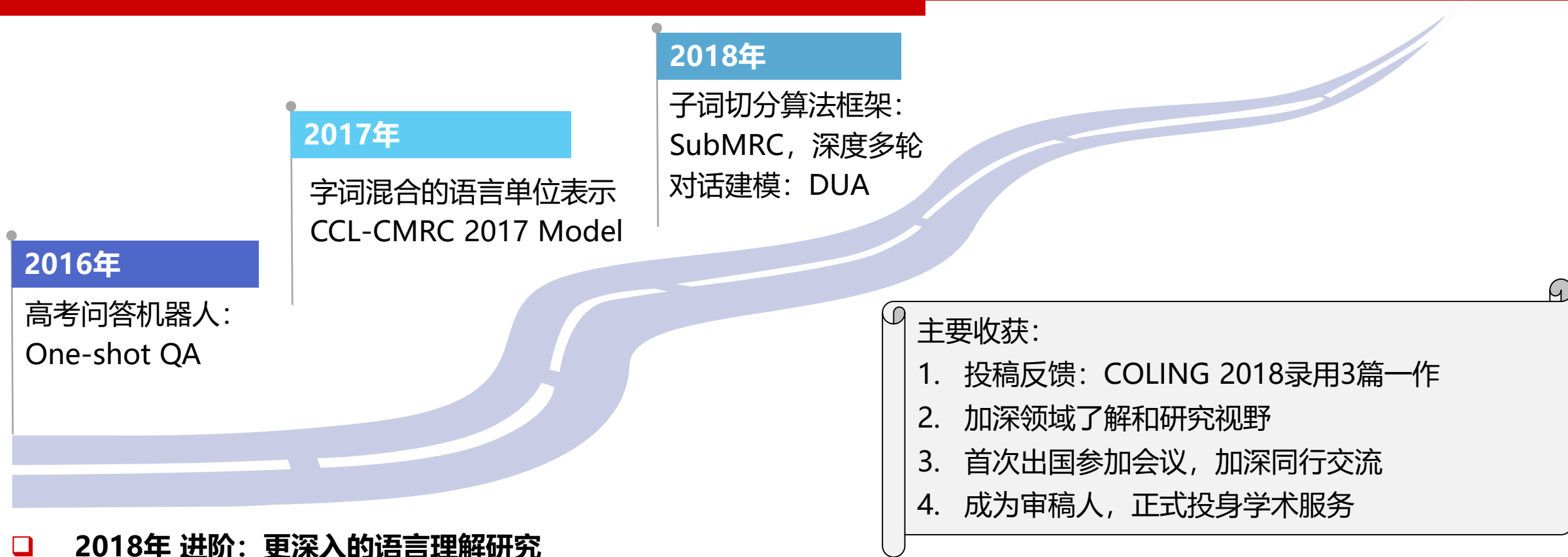
研究路线



❑ 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

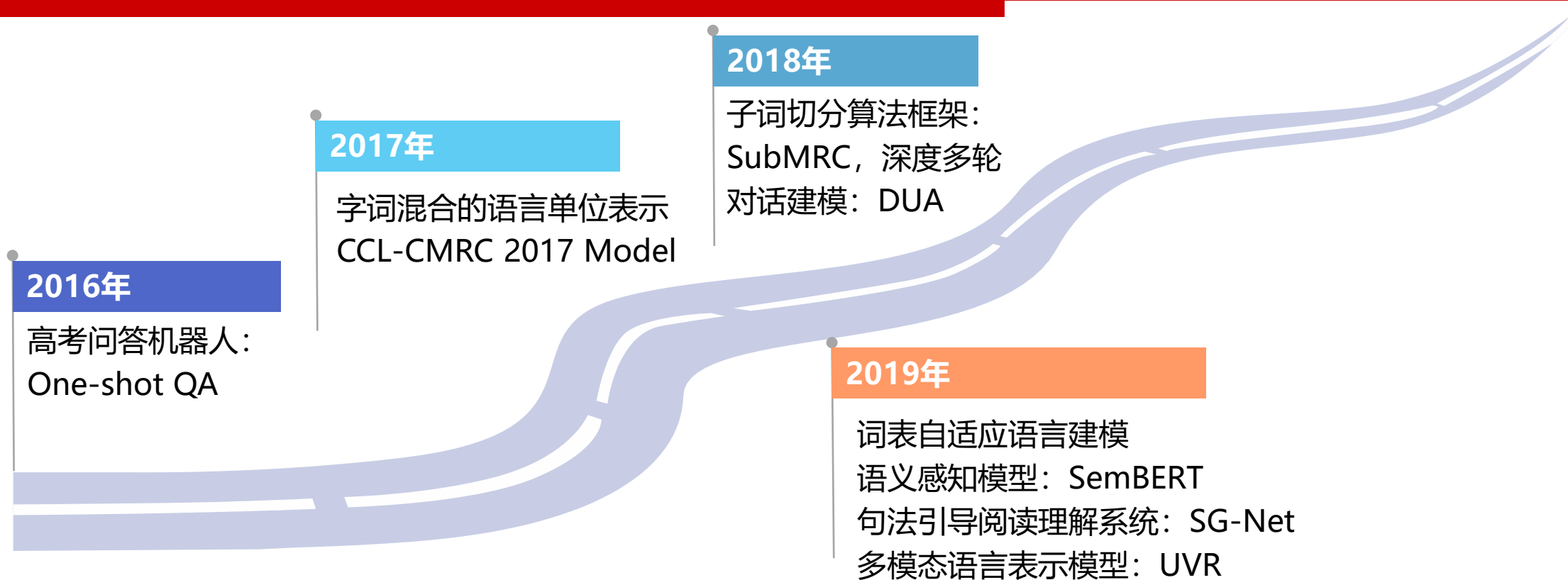
- 方法探索：寻找更灵活的语言单位粒度切分机制 (SubMRC, TASLP)
- 任务延申：探索多轮对话理解任务中的对话选择匹配问题 (DUA, COLING 2018)

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

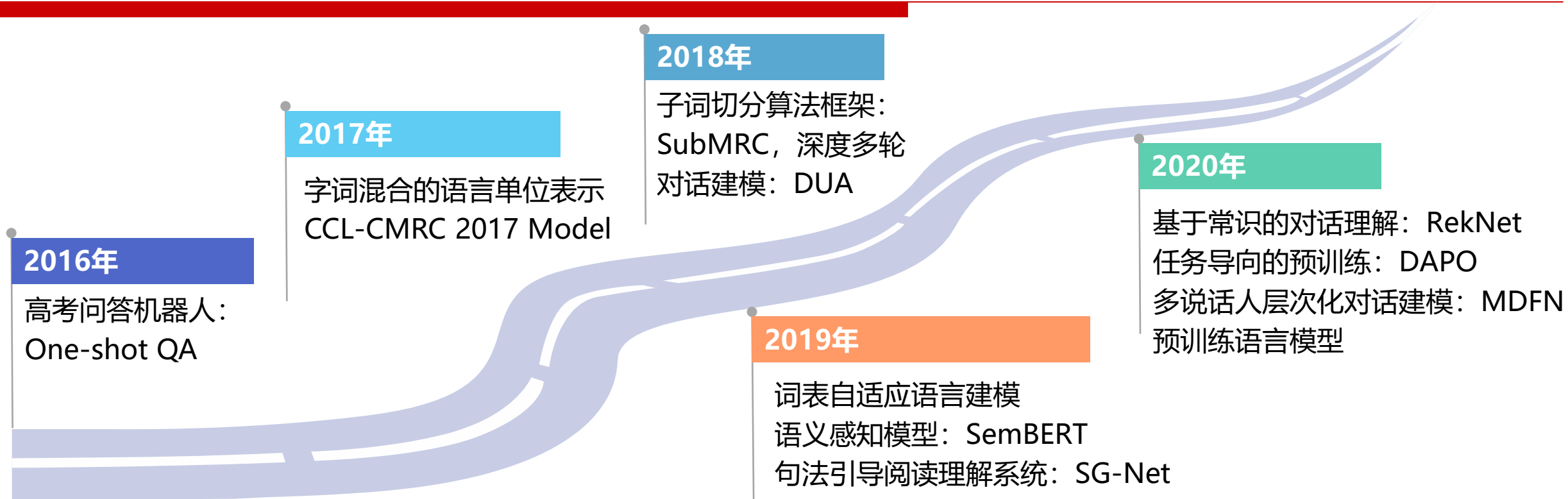
研究路线



□ 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, AACL 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, AACL 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



❑ 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

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

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：

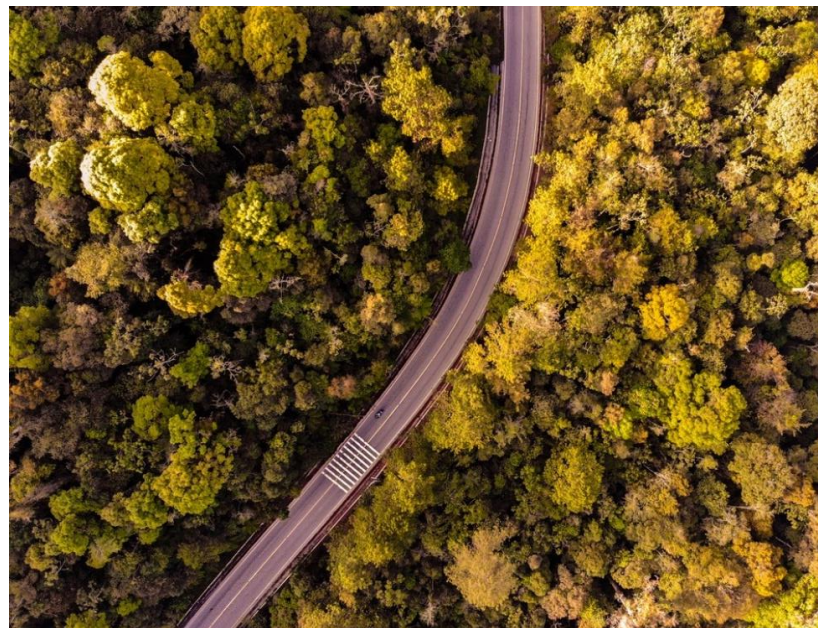
- 2017年 首届全国中文机器阅读理解评测（CMRC2017）单模型**第一名**
- 2019年 国际自然语言推理 SNLI 排行榜**第一名**
- 2019 国际权威机器阅读理解评测排行榜 SQuAD 2.0**第一名**
 - 首次以单模型超越**人类基准**
- 2019年国际大规模考试类阅读理解RACE**第一名**
- 2020年获得对话推理Mutual排行榜**第一名**

目录

- ❑ 研究历程
- ❑ 研究感悟

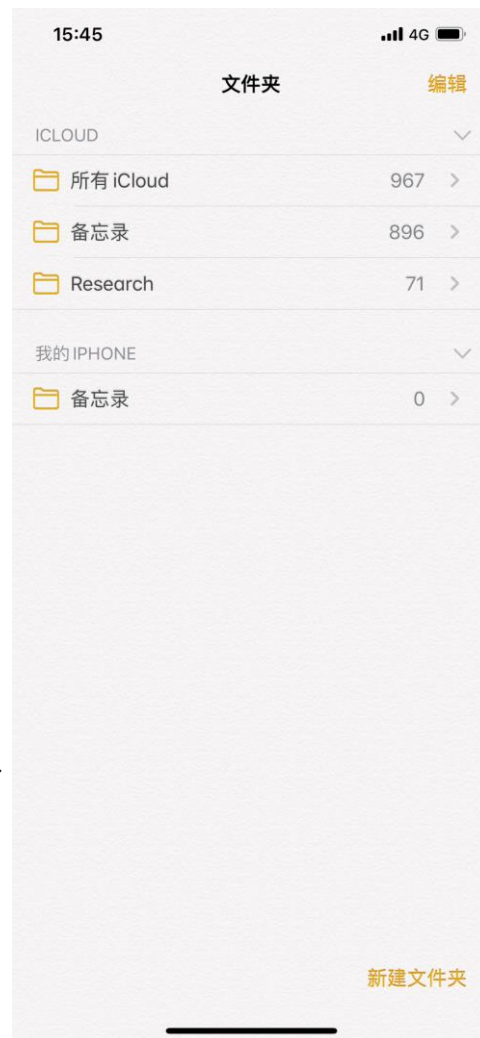
进入新领域

- 多读论文，总览全局
 - 视野：在一个森林，站的足够高，才不会一叶障目
- 从已有论文/算法寻找问题
 - 问题：有趣的问题、还未有效解决的问题
 - 方法：带着问题找方法
 - 实践：在实验中形成自己的认识
 - 验证：另辟蹊径，快速验证
- 尽快完成一套完整的科研训练，注重能力提升
 - 前期试错、积累经验、体验成败
 - 带着“问题”去读“文章”
 - 科研好比“自我模型”训练过程，根据反馈不断迭代进步



如何想到idea

- 收集、理解、思考
 - 每天看arXiv, 论文集、关注学者的主页、Twitter宣传等
 - 交流讨论, 相互学习
 - 随手记录, 经常总结
- 找到适合自己的“最佳独自思考环境”
- 积极探索有趣、有良好的应用场景、亟待突破的问题
- 不同阶段做不同挑战性的事情, 关注最核心的问题。以MRC为例:
 - 语言模型以前: 语言单位表示、篇章与问题匹配、答案抽取机制等
 - 语言模型之后: 语言模型预训练、多轮对话理解、常识知识型推理、开放域大规模阅读理解等



工作开展

- 站在巨人的肩膀上
 - 针对要解决的问题做好调研
 - 善于利用已有的轮子，学习优秀的代码/论文
- 做好版本控制，实验参数和结果记录
- 正视能力与技术发展的矛盾，放弃和选择同样重要
 - 不钻牛角尖（已有条件难以解决的问题，包括环境条件、个人能力条件等）
- 科研探索中的自我反思
 - 主动向导师、同学虚心请教，交流探讨，保持和谐的科研环境
 - 珍惜收到的反馈（如审稿人对论文的意见），从吐槽到认识工作的不足，并改进

应对遇到的困难

□ 潜在的困难：

- 论文看不懂、代码写不出、实验跑不通等

□ 了解自己的优势和缺陷，发挥优势，查漏补缺

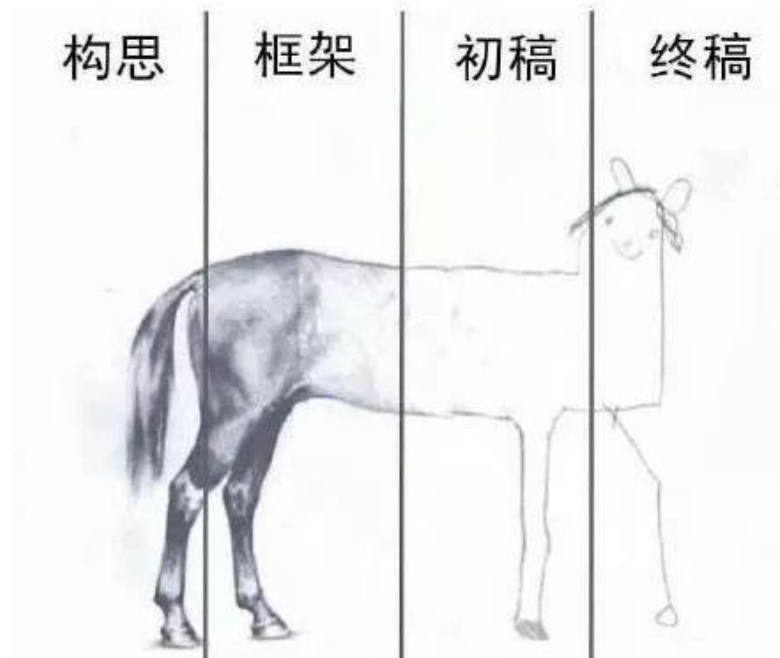
- 科研 = 理解 + 技术 + 展现
- 在解决问题的过程中提升科研能力和理解深度

□ 潜在的心理障碍

- 看论文的畏惧
- 对代码的畏惧
- 对写作的畏惧
- 对答辩的畏惧

□ 消除畏惧的最好方法就是面对它，跳出“舒适区”，主动寻找“正反馈”，不断提升：

- 如写作：从workshop paper -> conference paper -> standard journal paper -> long survey paper



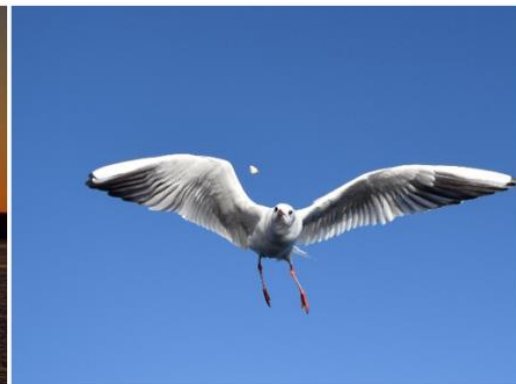
如何应对精神压力

□ 压力来源:

- 科研不顺利、同行竞争大、未来不确定性等
- 急于求成、盲目攀比、好高骛远

□ 排解方法:

- 不要盲目攀比，做好自己，看到自己的进步
- 积极交流讨论，保持心理健康
- 培养自己的爱好，锻炼身体，如摄影、乐器、读书、电子游戏（适当）、游泳等
- 世界那么大不如去看看



谢谢大家！

Homepage: <http://bcmi.sjtu.edu.cn/~zhangzs>

E-mail: zhangzs@sjtu.edu.cn