Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

文本生成中的幻觉

硕士研究生 杨宗源

2023年08月20日

问题回溯

- 总结反思
  - 演讲表达能力较差，报告整体流畅度不高
  - 内容深度不足

- 相关内容
  - 张凌浩《基于图结构处理的文本生成》——2022.02.27
  - 高依萌《预训练语言模型GPT-3》——2021.02.07

预期收获

- 预期收获
  - 1.了解文本生成基本原理
  - 2.理解文本生成中幻觉问题的研究脉络
  - 3.理解将外部知识引入文本生成的方法和原理
  - 4.了解幻觉的评价方法

内容提要

- 背景简介
- 基础概念
  - 文本生成
  - 幻觉
  - 评价指标
  - 常见解决方法
- 算法原理
  - KIDReview
  - CaPE
- 应用总结
- 前沿发展
- 参考文献

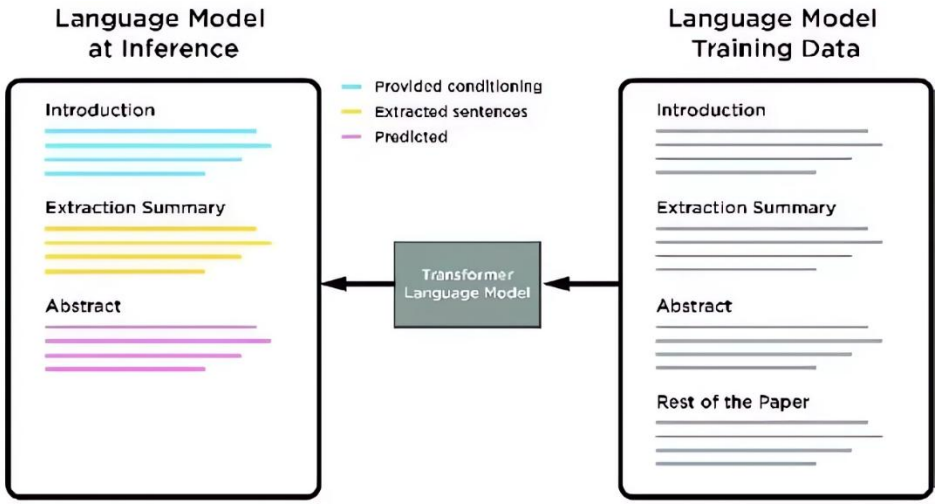
研究背景与意义

- 文本生成快速发展
  - 2022年12月，大型预训练模型ChatGPT横空出世
  - 短短两个月活跃用户就超过一亿
  - 各类生成式大模型受到人们广泛关注
  - 文本生成的热度不断上升

- 引发问题
  - 无中生有、臆造事实、缺乏信息量
  - 极大程度限制了文本生成模型的进一步应用
  - Stack Overflow禁止ChatGPT在网站中使用
  - 《互联网信息服务深度合成管理规定》对深度合成技术细化监管要求

基本概念 文本生成

- 文本生成
  - 以一种或多种自然语言自动生成人类可理解文本的过程
- 常见任务类型
  - 机器翻译（Machine Translation）
  - 抽象摘要（Abstract Summarization）
  - 对话生成（Dialogue Generatioin）
  - 代码生成（Code Generatioin）
  - ……
- 发展过程



基本概念 文本生成基本原理

- 自回归文本生成的原理
  - 文本条件概率计算公式

$$P_M(y|x) = \prod_{i=1}^{n} P_M(y_i|y_{<i}, x)$$

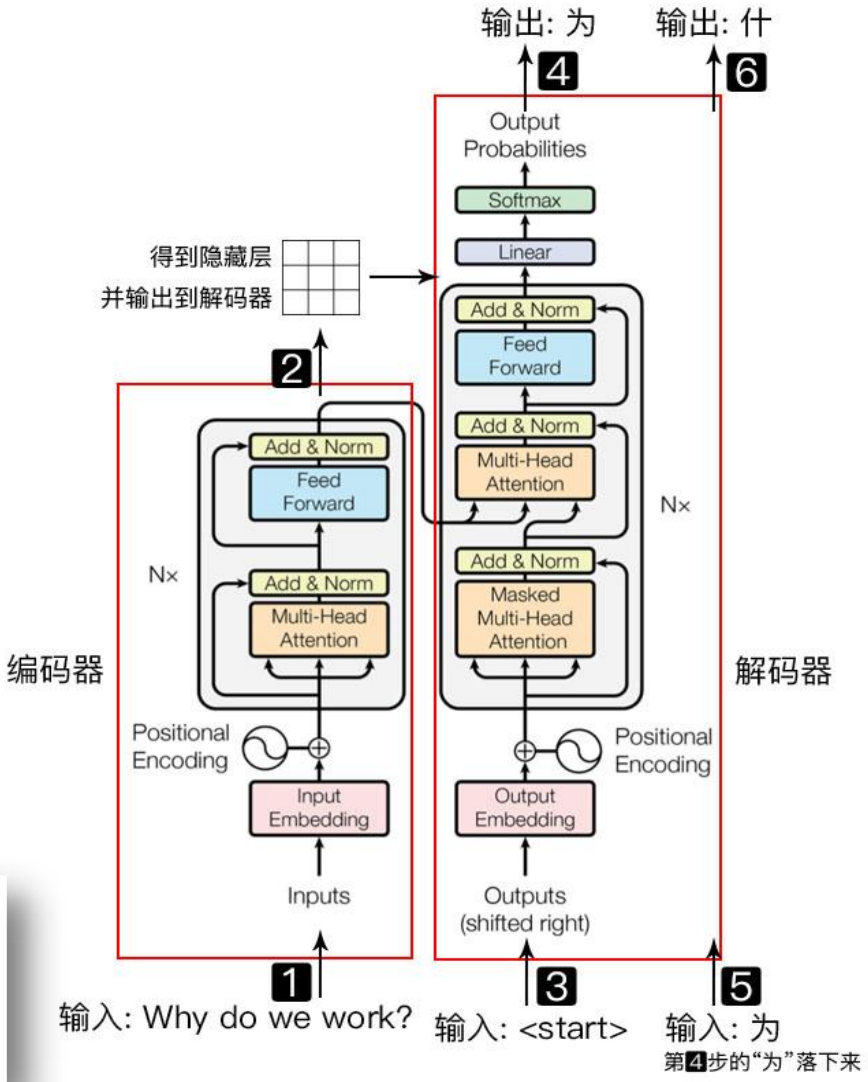
  - 最大似然法计算Loss

$$L(\theta) = \text{argmax} \sum_{i=1}^{n} logP\ (y_i|y_{<i}, x; \theta)$$

该过程可能引发幻觉！

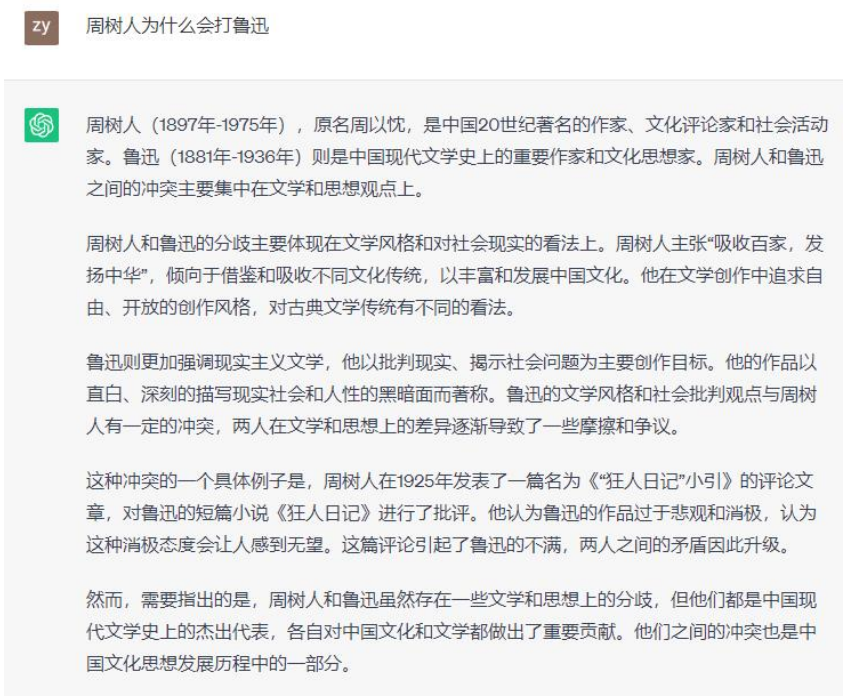
Question: definition of sadduction

Answer: Sadducees. noun. **The definition of a sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power.**1 a member of a Jewish sect that was active during the Second Temple.



基本概念 幻觉

- 基本定义
  - 生成的无意义或不忠实于所提供的源内容的文本
- 类别
  - 内在幻觉
    - 生成的输出与源内容相矛盾
  - 外在幻觉
    - 生成的输出无法从源内容中得到验证
- 不同任务对幻觉的容忍度与定义也有所差别

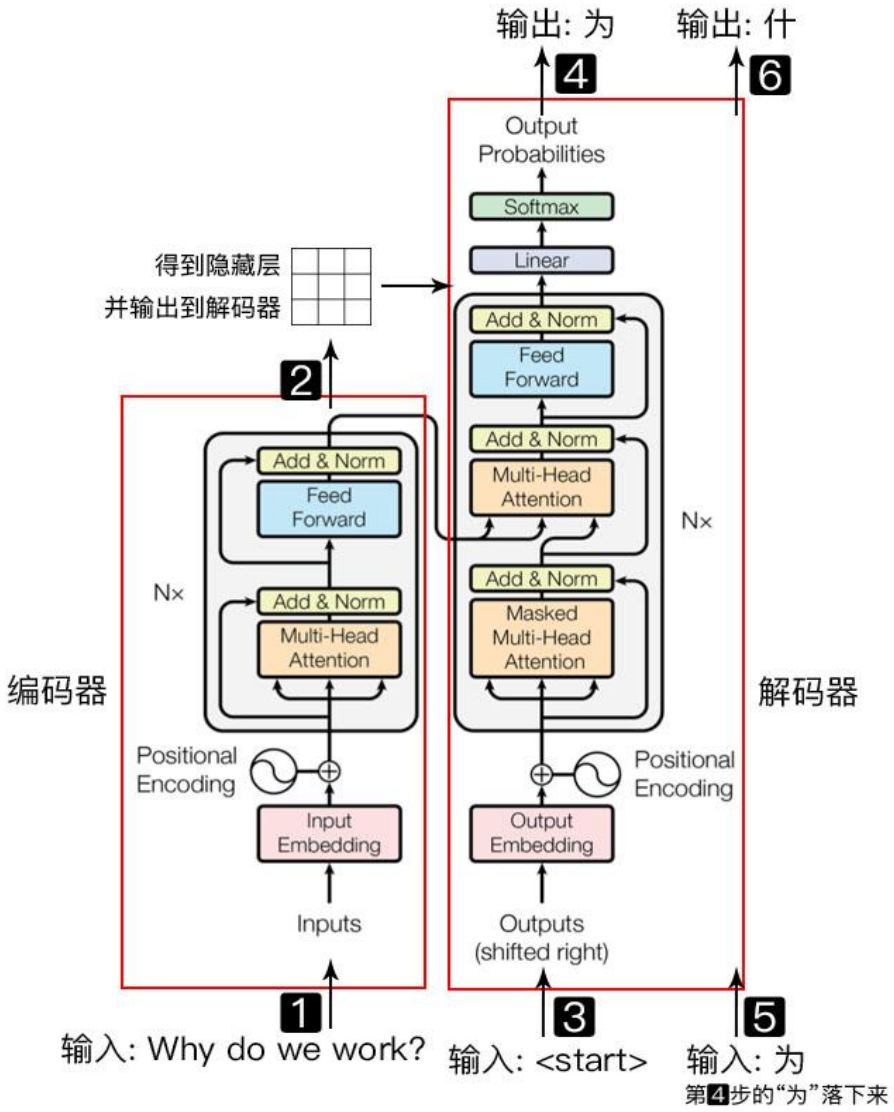


	TEAM	CITY	WIN	LOSS	PTS	FG_PCT	BLK
Intrinsic	Rockets	Houston	18	5	108	44	7
Extrinsic	Nuggets	Denver	10	13	96	38	7

The Houston Rockets **(18-4)** defeated the Denver Nuggets (10-13) 108-96 on Saturday.
Houston has won two straight games and six of their last seven.

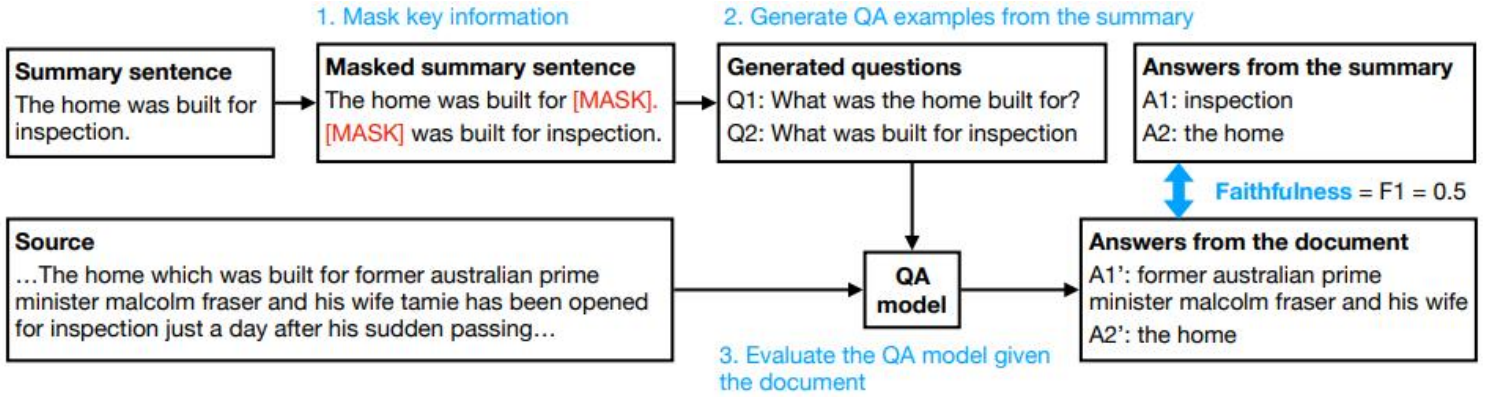
基本概念 产生幻觉的原因

- 数据
  - 数据收集过程不合理
  - 无法保证事实知识一致性
- 训练推理过程
  - 输入文本的表示不完备
    - 编码器学习到训练数据的错误相关性
  - 解码错误
    - 解码器过度关注输入源的错误部分
  - 误差累积
    - seq2seq的逐个生成方式会产生误差累积
  - 参数知识偏差
    - 预训练模型倾向使用参数知识，忽视微调信息

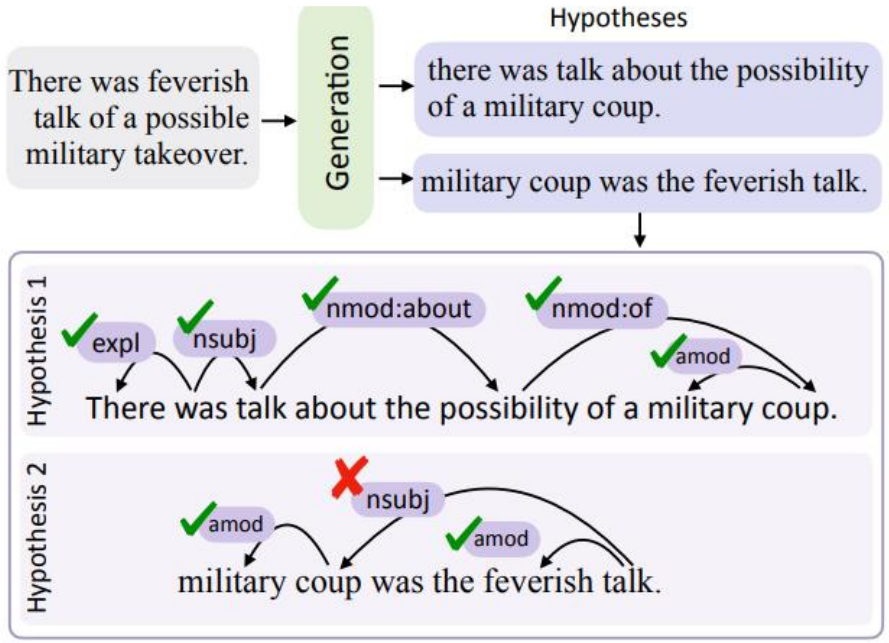


基本概念 幻觉评价指标

- 统计指标
  - N-gram，Rouge，BLEU
- 基于模型度量
  - 基于问答
  - 基于信息抽取
  - 基于自然语言推理
  - 基于分类模型
  - 基于预训练模型
- 人类评估
  - 对幻觉水平评分
  - 将输出文本与源文本进行比较



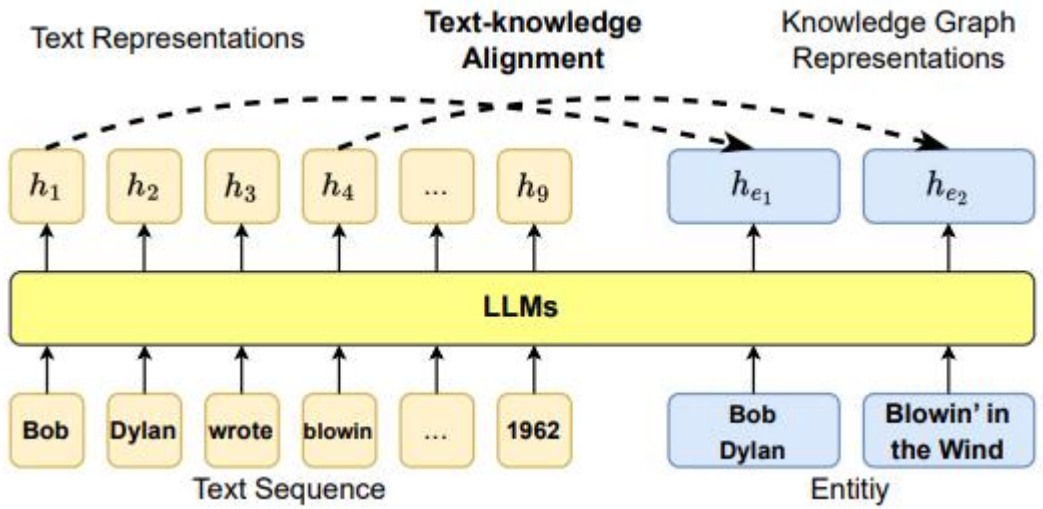
问答

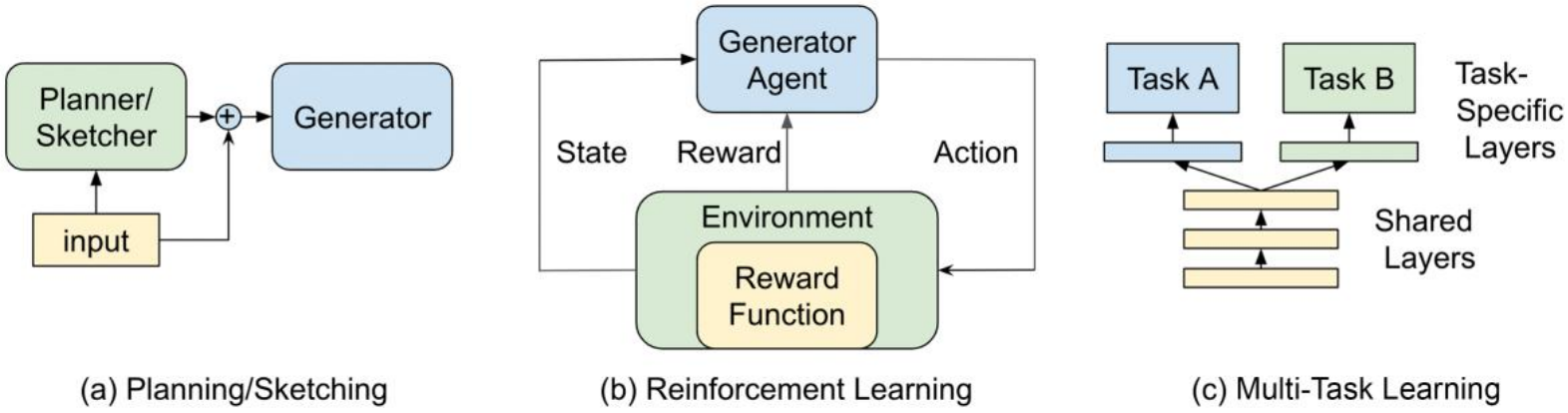


信息抽取

基本概念 缓解幻觉的方法

- 数据相关方法
  - 构建可信数据集
  - 自动清理数据
  - 信息增强
- 建模推理方法
  - 架构控制
  - 训练方式优化
  - 多任务学习
  - 可控生成
  - 其他
- 后处理

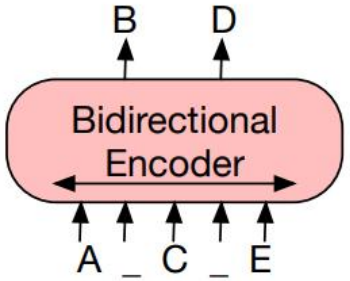


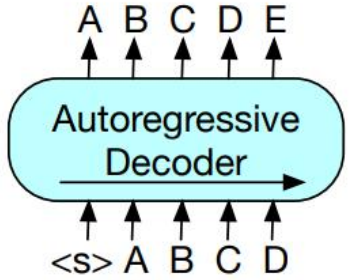


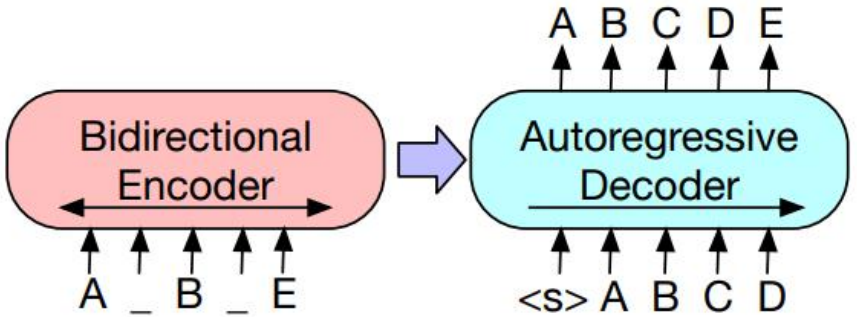
(a) Planning/Sketching (b) Reinforcement Learning (c) Multi-Task Learning

基本概念 BART

- BERT
  - 仅使用Encoder
  - 将源文本进行自关注获得文档中每个词的表示
- GPT
  - 仅使用Decoder
  - 以左侧上下文为条件进行文本生成
- BART
  - 同时使用Encoder和Decoder
  - 使用双向模型对损坏文档进行编码，然后使用自回归解码器生成文本
  - 预训练时使用多种加噪方式训练模型恢复重建能力







【 KIDReview 】

KIDReview TIPO PCDL

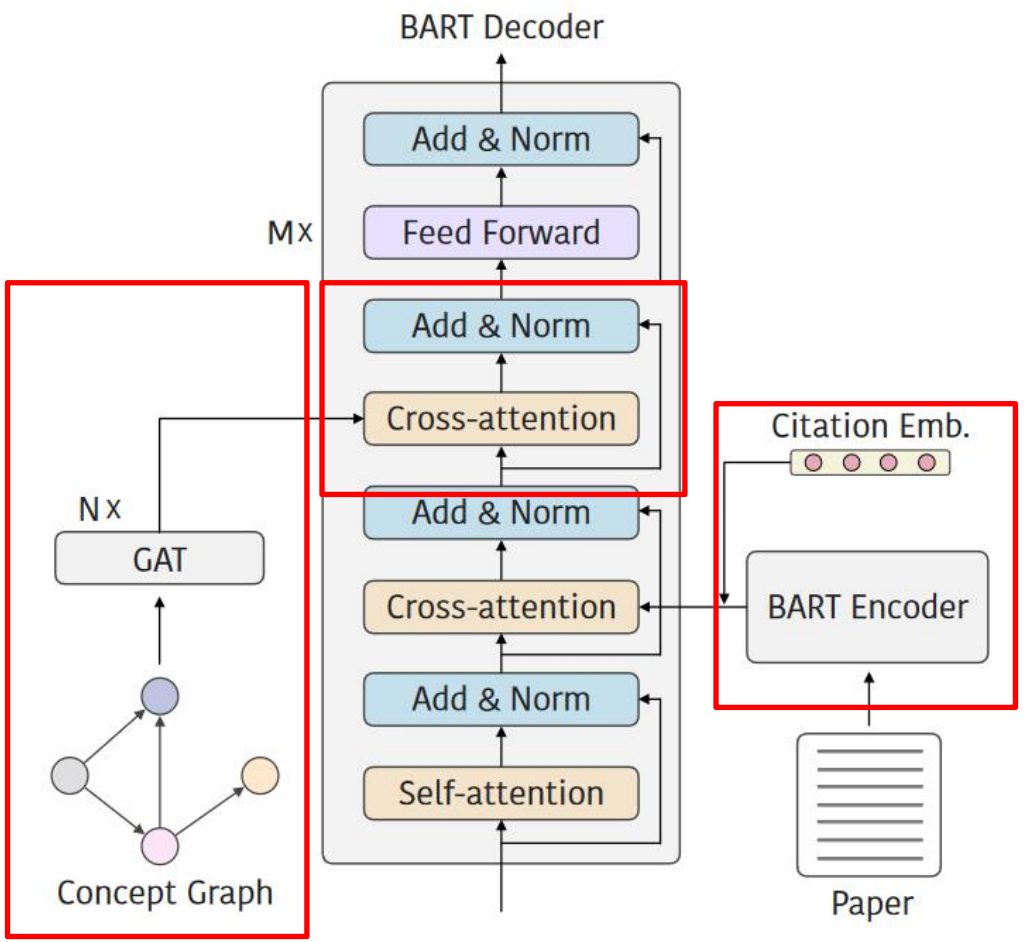
T	目标	根据论文生成论文评审意见
I	输入	论文文本（ICLR*5.1k篇, NIPS*3.6k篇） 真实评审意见（ICLR*15.7k篇, NIPS*12.3k篇） 引文论文数据集（8110万篇学术论文）
P	处理	1. 使用oracle文本预训练BART模型 2. 使用交叉熵提取关键句子 3. 提取概念图训练知识嵌入，将实体知识引入BART模型 4. 构建引文图训练引文嵌入，将外部引文知识引入BART模型
O	输出	生成的论文评审意见文本*8.7k篇

P	问题	生成的意见文本容易出现非真实的幻觉
C	条件	需要预先构建关键词表以提取关键句子
D	难点	如何将外部知识引入生成模型
L	水平	CCF-A AAAI 2022

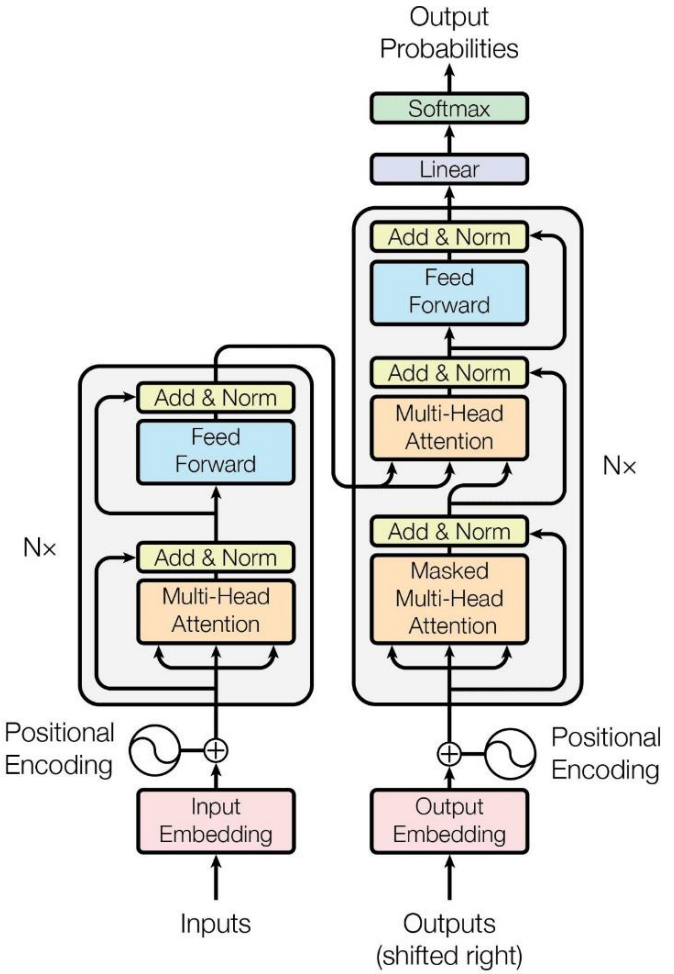
KIDReview 框架与原理

- 算法原理图

KIDReview



BART



- 引文图构建
  - 使用大量计算机领域论文构建引文无向图
  - 计算引文嵌入
    - 对于每条无向边，临近节点联合概率为

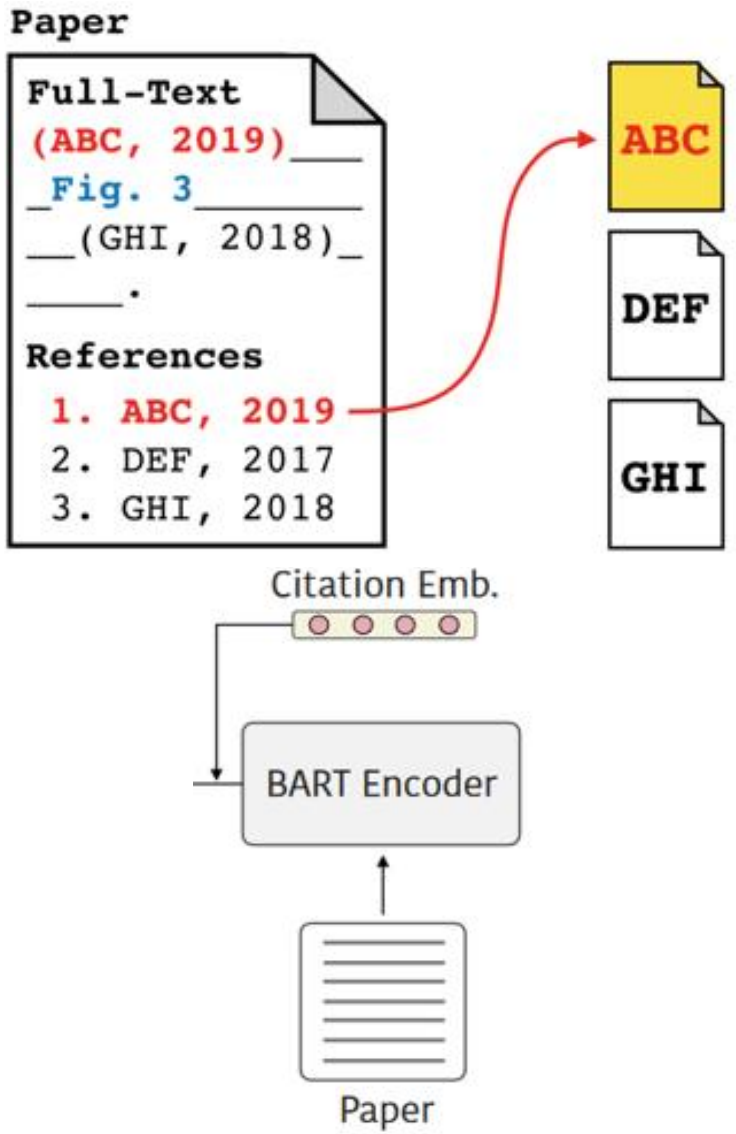
$$p_1(v_i, v_j) = \frac{1}{1 + \exp(-\vec{u}_i^T * \vec{u}_j^T)}$$

    - 优化目标函数为

$$O_1 = -\sum_{(i,j)\in E} w_{ij} log p_1(v_i, v_j)$$

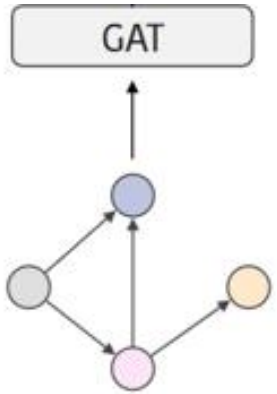
  - 论文表示结合引文嵌入

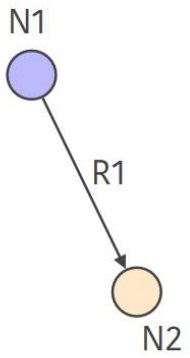
$$x = concat(W_c c, x')$$



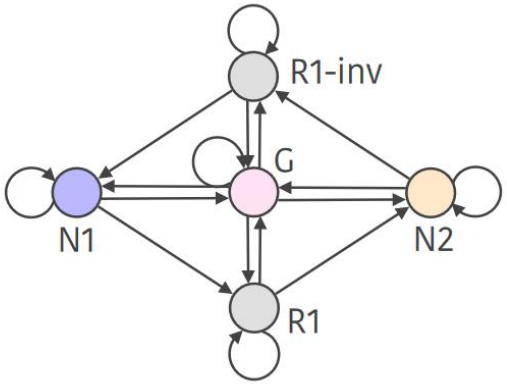
- 概念图构建
  - 使用SciERC提取论文中的实体和关系
    - 实体：任务、材料、方法、度量、通用、其他
    - 关系：部分、用于、比较、特征、同义词、评估、结合
  - 使用Knocel-Kedziorski法重构概念图
    - 原始实体节点和关系边通常不能形成连通图
    - 难以直接应用图神经网络进行训练
    - N为实体节点，R为关系节点，
    - R-inv为逆关系节点，G为全局节点
  - 使用GAT训练概念图

$$e_i' = LN(FFN(\widetilde{e_i}) + \widetilde{e_i})$$





(a) Before transformation.



(b) After transformation.

- 知识引入
  - 解码器层添加一层交叉注意力模块
  - 根据前一层表示计算当前层表示

$$\tilde{y}^{l+1} = LN(y^l + SelfAttn(y^l))$$

  - 引入结合引文嵌入的论文表示

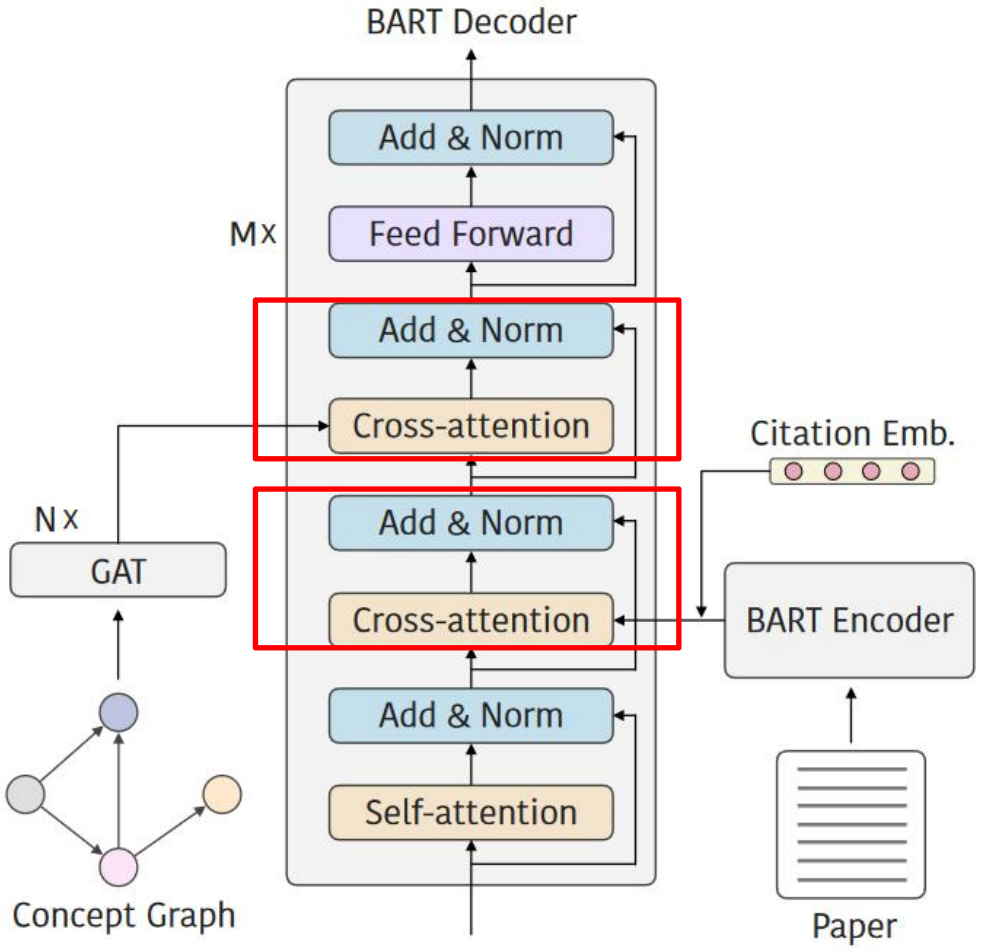
$$\tilde{y}^{l+1} = LN(\tilde{y}^{l+1} + CrossAttn(\tilde{y}^{l+1}, x))$$

  - 引入实体表示

$$\tilde{y}^{l+1} = LN(\tilde{y}^{l+1} + CrossAttn(\tilde{y}^{l+1}, e))$$

  - 维度转化调整

$$\tilde{y}^{l+1} = LN(\tilde{y}^{l+1} + FFN(\tilde{y}^{l+1}))$$



KIDReview 数据处理方法

- Oracle文本提取
  - 提取最高平均Rouge分数的文本

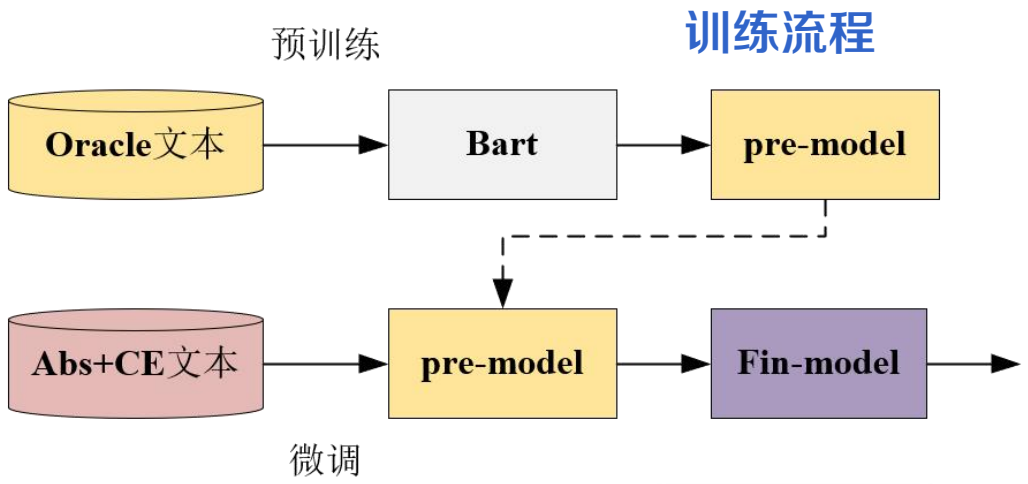
$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_N \in S} Count_{match}(gram_N)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_N \in S} Count(gram_N)}$$

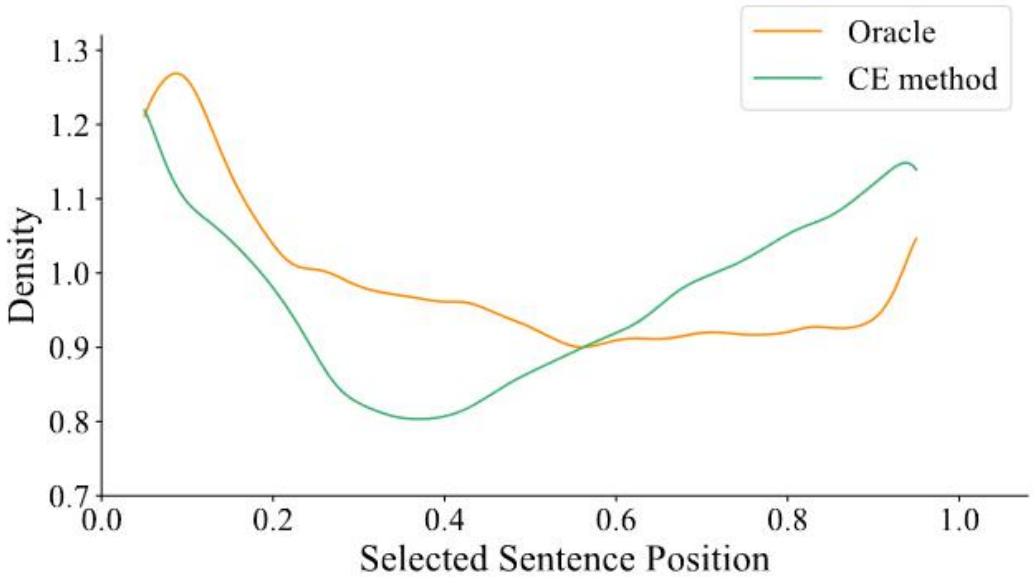
- 交叉熵文本提取
  - 预定义关键词提取句子
  - 将句子提取问题转化为组合优化问题

$$p_s(w) = \frac{count(w)}{Len(S)}$$

$$R(S) = -\sum_{w \in S} p_S(w) log p_S(w)$$

  - 结合论文摘要文本和交叉熵提取文本





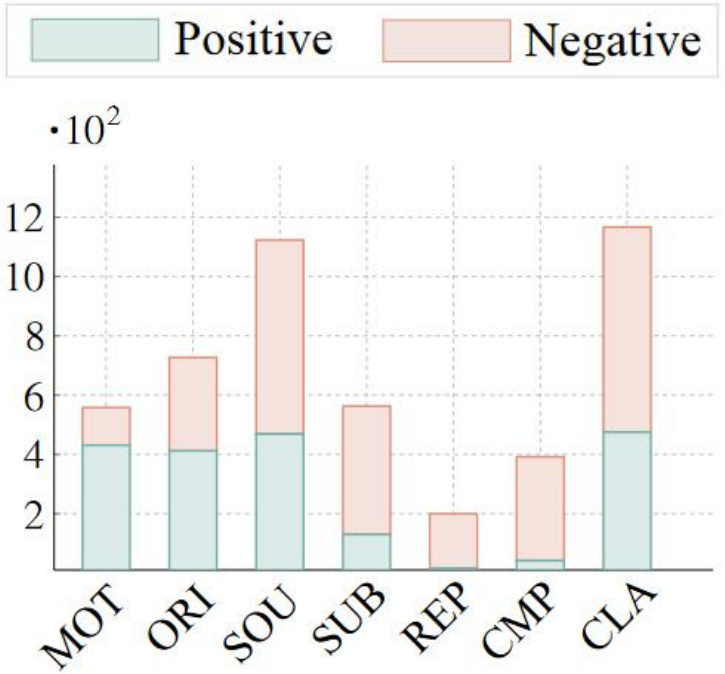
KIDReview 实验设计

- 数据集
  - ASAP-Review数据集
    - 包括8.8k英文论文及28k评审意见
    - 每个意见包括7种方面的倾向注释
  - S2ORC数据集
    - 包括81.1M英文论文及其引用关系

- 实验设计
  - S2ORC数据集训练引文嵌入
  - ASAP-Review数据集训练概念图及模型微调

	Accept	Reject	# of Reviews
ICLR	1,859	3,333	15,728
NeurIPS	3,685	0	12,391

Table 1: Basic statistics of ASAP-Review dataset.



KIDReview 实验结果

- 客观评价
  - 推荐准确率R_{ACC}
    - 根据生成的评审意见判断论文是否被接受与原始判断的一致性

$$R_{ACC} = DEC(paper) \times REC(review)$$

  - 方面覆盖率A_{COV}
    - 在预定义的方面类型中覆盖了多少方面
  - 方面召回率A_{REC}
    - 在生成的文本中覆盖了多少源评审意见中的方面

通过训练的分类模型进行评估

	Pre.	Knowledge	RACC	ACOV	AREC
Human	–	–	49.25	50.83	58.35
Oracle	–	vanilla	2.40	67.51	65.28
	–	+ citation	**10.06**	68.66	**67.48**
	–	+ concept	6.86	**71.77**	65.74
	–	+ cit.& con.	5.03	67.67	64.09
CE	✗	vanilla	13.94	62.64	60.73
	✓	vanilla	11.43	**67.39**	**62.56**
	✓	+ citation	12.80	66.90	62.49
	✓	+ concept	12.11	62.01	60.85
	✓	+ cit. & con.	**23.31**	61.00	61.99
Abs.+CE	✗	vanilla	15.54	55.37	58.31
	✓	vanilla	17.03	63.47	63.00
	✓	+ citation	21.14	**64.69**	**63.53**
	✓	+ concept	18.06	60.64	59.80
	✓	+ cit. & con.	**25.03**	58.46	60.90

KIDReview 实验结果

- 客观评价结论
  - Oracle预训练使模型更关注于训练数据，增强微调效果
  - 加入概念和引文知识后R_{ACC}明显提高，生成意见的判断更准确，幻觉更少
  - Oracle微调有最高的A_{COV}和A_{REC}，说明处理长文本输入时高效的内容选择策略仍有价值

	Pre.	Knowledge	RACC	ACOV	AREC
Human	–	–	49.25	50.83	58.35
Oracle	–	vanilla	2.40	67.51	65.28
	–	+ citation	**10.06**	68.66	**67.48**
	–	+ concept	6.86	**71.77**	65.74
	–	+ cit.& con.	5.03	67.67	64.09
CE	✗	vanilla	13.94	62.64	60.73
	✓	vanilla	11.43	**67.39**	**62.56**
	✓	+ citation	12.80	66.90	62.49
	✓	+ concept	12.11	62.01	60.85
	✓	+ cit. & con.	**23.31**	61.00	61.99
Abs.+CE	✗	vanilla	15.54	55.37	58.31
	✓	vanilla	17.03	63.47	63.00
	✓	+ citation	21.14	**64.69**	**63.53**
	✓	+ concept	18.06	60.64	59.80
	✓	+ cit. & con.	**25.03**	58.46	60.90

KIDReview 实验结果

- 主观评价
  - 准确性评价
    - 选取40份未纳入训练集的论文进行意见生成
    - 得分为1表示同意，0为不同意，0.5为部分同意
  - 建设性评价
    - 每篇意见的建设性程度进行配对排名
    - 计算关键短语在生成文本中的出现次数

	vanilla	vanilla(Pre.)	+cit.&con. (Pre.)
vanilla	×	47.73	45.45
vanilla(Pre.)	**52.27**	×	42.86
+cit.&con. (Pre.)	**54.55**	**57.14**	×

- 主观评价结论
  - 添加引文/概念知识，建设性与准确性均高于基线

	Vanilla	Vanilla (Pre.)	+ cit.&con. (Pre.)
SACC	39/40	40/40	39.5/40

	Vanilla	+ cit.	+ con.
for example	615	616	**680**
e.g.	740	757	741
such as	255	261	**282**
for instance	294	294	**394**
should compare	90	115	**170**
questions	22	25	**38**
?	378	347	**411**

KIDReview 优劣分析

- 优势
  - 将多种知识引入预训练模型，提升生成文本的效果
  - 使用Oracle预训练策略，避免预训练模型过度关注参数知识

- 劣势
  - 交叉熵文本提取方法需要手动设置关键词，缺乏泛化性
  - 无法主动判断幻觉存在，生成的文本仍然存在外部幻觉

【CaPE】

CaPE TIPO PCDL

T	目标	生成文档摘要
I	输入	文档原始文本（BBC新闻及摘要*22万条，BBC+每日邮报新闻及摘要*32万条）
P	处理	1. 识别幻觉样本 2. 训练基本模型 3. 分别使用清洗前后的数据训练专家模型和反专家模型 4. 集成三个模型参数获得对比参数集成模型
O	输出	生成的新闻文档摘要*54万条

P	问题	消除生成文本中存在的幻觉
C	条件	提前判断幻觉样本构建干净数据集
D	难点	如何确定幻觉样本的评判标准
L	水平	CCF-A ACL 2023

CaPE 框架与原理

- 幻觉样本判断
  - 实体令牌精度
    - 在摘要中出现实体可在源文本中找到的百分比

$$E - P_{src} = \frac{N(h \cap s)}{N(h)}$$

  - 依赖弧蕴含误差（DAE error）
    - 使用预训练模型提取实体和关系构建依赖弧

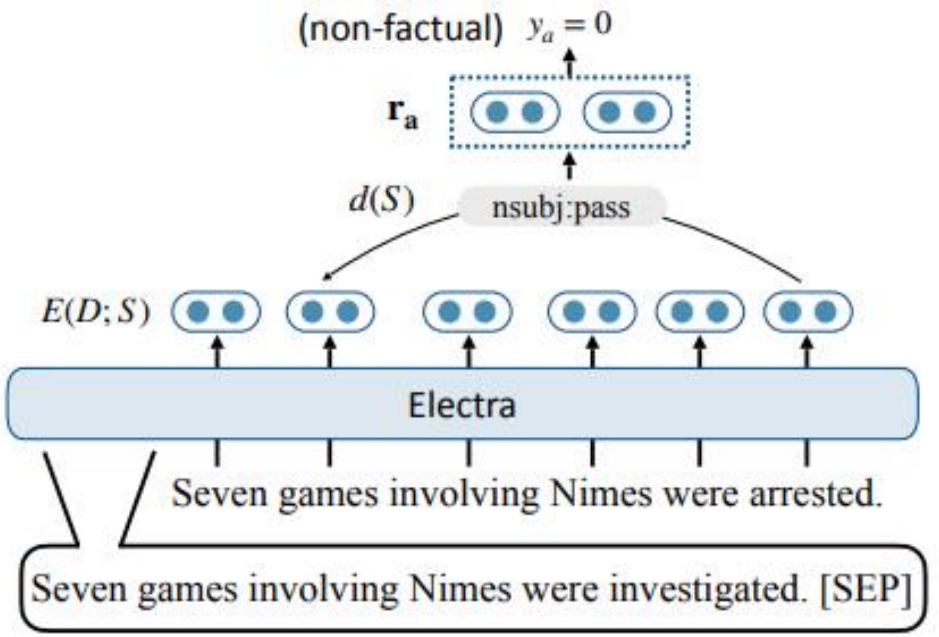
$$r_a = [E(x;h)_{a_h}; E(x;h)_{a_c}; E(a_d)]$$

    - 在摘要中而不在源文档中依赖弧的数量

$$DAE_{err} = N_r(h) - N_r(h \cap s)$$

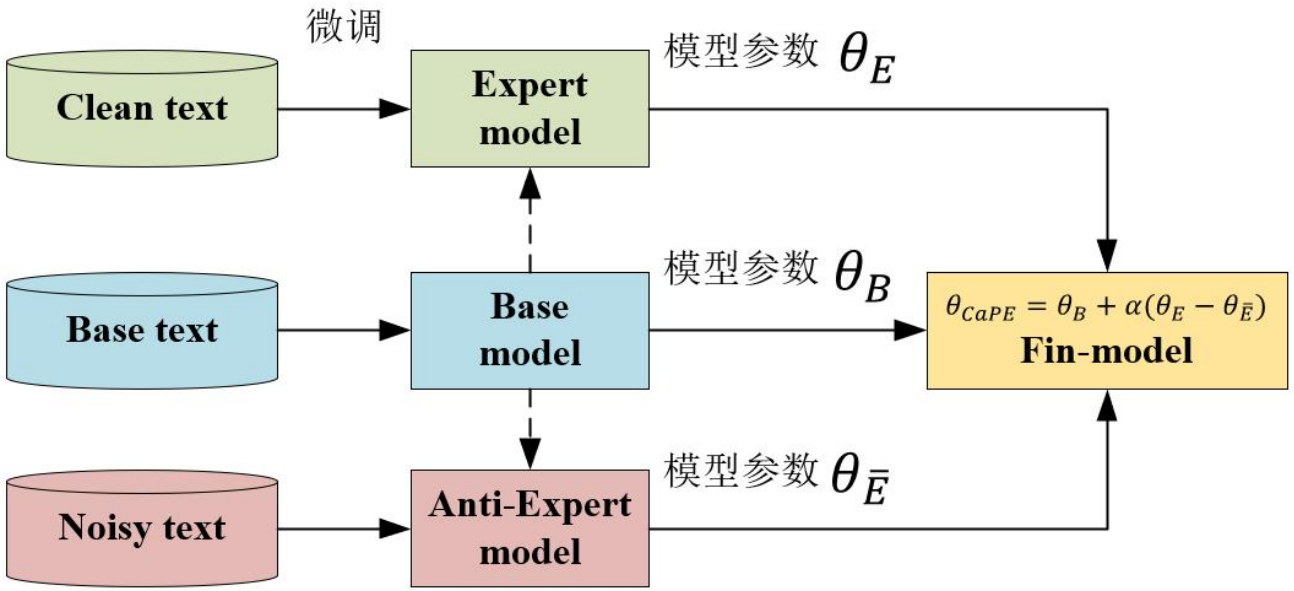
  - 分别设定阈值区分干净样本和幻觉样本

训练模型提取实体/依赖弧



CaPE 框架与原理

- 对比参数集成模型构建
  - 使用原始数据训练得到基础模型
  - 使用干净/幻觉数据微调基础模型得到专家/反专家模型
  - 结合三种模型参数构造集成模型



Algorithm 1 CaPE for Summarization

Require: Training Data D_T, Measure of hallucination M_H

1: Train θ_B on D_T
2: $D_{clean} \leftarrow$ SELECTCLEAN (D_T, M_H)
3: $D_{noisy} \leftarrow$ SELECTNOISY (D_T, M_H)
4: $\theta_E \leftarrow$ Fine-tune θ_B on D_{clean}
5: $\theta_{\bar{E}} \leftarrow$ Fine-tune θ_B on D_{noisy}
6: $\theta_{CaPE} \leftarrow \theta_B + \alpha(\theta_E - \theta_{\bar{E}})$
7: **return** θ_{CaPE}

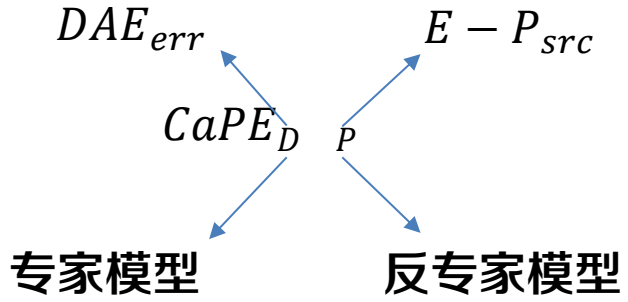
CaPE 实验设计

- 评价指标

指标类型	指标解释	指标符号
事实一致性	生成摘要中依赖弧占源文档中依赖弧的百分比	$D_{arc} = \frac{N_r(h)}{N_r(s)}$
	无错误依赖弧的摘要占所有生成摘要的百分比	$D_{sum} = \frac{N_s\ (rsum)}{N_s\ (sum)}$
	源文档实体精确率	$E - P_{src} = \frac{N(h \cap s)}{N(h)}$
	源摘要实体召回率	$E - R_{ref} = \frac{N(h \cap r)}{N(h)}$
	基于QA模型的事实一致性度量	$QEval$、$QAFactEval$
	基于MNLI数据集的RoBERT模型分数	$MNLI$
语义一致性	BertScore的精确率和召回率	$BS - P(R)$
	Rouge-N和Rouge-L分数	$Rouge - 1 \setminus 2 \setminus \mathrm{L}$
时间	训练用时 / 测试用时	$TT \setminus \mathrm{IT}$

CaPE 实验结果

- 数据集
  - XSUM、CNN/DM
- 对比实验
  - BARTsum、Ensemble、
  - PP、PP-Clean、 PP-CC
  - 四种CaPE模型
- 下标解释



Data	Exp_{DAE}	$Anti_{DAE}$
XSUM	39009 (19.1%)	7962 (3.9%)
CNN/DM	39643 (13.8%)	8786 (3.1%)
	Exp_{E-P}	$Anti_{E-P}$
XSUM	50270 (24.6%)	26208 (12.8%)
CNN/DM	152418 (53.0%)	31727 (11.1%)

Model	D_{arc}	D_{sum}	E-P_{src}	E-R_{ref}	QEval	BS-P	BS-R	R1	R2	RL	TT	IT
XSUM												
Base	76.16	34.75	63.82	53.66	36.54	88.93	79.86	45.34	22.21	37.13	1x	1x
Ensemble	75.22	33.48	62.63	**54.23**	36.37	88.82	79.86	45.27	22.28	37.09	1.2x	1x
PP	75.65	33.67	62.36	53.93	36.37	88.88	79.84	45.34	**22.30**	37.18	2-3x	2x
PP-Clean	79.41	40.09	**72.98**	45.72	37.01	89.09	79.84	43.82	20.40	35.89	1.5x	2x
PP-CC	76.88	35.99	66.06	52.23	36.62	88.95	79.85	45.03	21.87	36.89	-	2x
$CaPE_{DD}$	78.51	39.36	65.61*	52.91	36.90	89.08	79.81	45.33	22.29	**37.27**	1.05x	1x
$CaPE_{PP}$	78.46	39.13	69.12	53.36	37.09	89.07	**79.89**	45.16	21.91	36.94	1.08x	1x
$CaPE_{DP}$	**79.61**	**40.55**	68.24	53.91	**37.22**	**89.15**	**79.89**	45.14	21.97	36.92	1.07x	1x
$CaPE_{PD}$	77.91	38.40	66.12*	52.77	36.84	89.05	79.81	**45.35**	22.25	37.17	1.06x	1x
$CaPE_{DP*}$	83.87	48.78	74.30	52.34	38.05	89.41	79.93	43.56	20.39	35.46	1.07x	1x
CNN/DM												
Base	96.26	75.0	98.44	58.92	59.24	93.26	82.62	44.05	21.07	40.86	1x	1x
Ensemble	95.19	67.44	97.72	**61.93**	59.51	93.06	**82.91**	**44.28**	**21.23**	**40.88**	1.2x	1x
PP	96.14	74.70	98.26	58.40	59.15	93.23	82.58	43.95	20.94	40.76	2-3x	2x
PP-Clean	96.17	74.77	98.63	58.20	59.16	93.23	82.59	43.92	20.92	40.74	2x	2x
PP-CC	95.72	72.63	98.52	58.57	59.11	93.22	82.61	43.97	20.98	40.79	-	2x
$CaPE_{DD}$	**98.23**	**86.54**	98.90	58.35	**60.10**	**93.80**	82.84	43.75	20.79	40.44	1.04x	1x
$CaPE_{PP}$	97.17	80.46	**99.16**	58.66	59.65	93.52	82.71	43.62	20.72	40.33	1.14x	1x
$CaPE_{DP}$	97.59	83.04	98.86	58.86	59.70	93.56	82.78	43.71	20.80	40.42	1.06x	1x
$CaPE_{PD}$	96.97	79.39	98.66	58.60	59.61	93.44	82.68	44.05	21.07	40.83	1.11x	1x

CaPE 实验结果

- 事实一致性分析
  - CaPE在大部分事实一致性（幻觉）指标上有明显提升
  - 通过调整超参数，模型可以在各类指标上进一步提升
  - 集成模型存在信息损失，导致源摘要实体召回率$E-R_{ref}$下降

Model	D_{arc}	D_{sum}	E-P_{src}	E-R_{ref}	QEval
					XSU
Base	76.16	34.75	63.82	53.66	36.54
Ensemble	75.22	33.48	62.63	**54.23**	36.37
PP	75.65	33.67	62.36	53.93	36.37
PP-Clean	79.41	40.09	**72.98**	45.72	37.01
PP-CC	76.88	35.99	66.06	52.23	36.62
$CaPE_{DD}$	78.51	39.36	65.61*	52.91	36.90
$CaPE_{PP}$	78.46	39.13	69.12	53.36	37.09
$CaPE_{DP}$	**79.61**	**40.55**	68.24	53.91	**37.22**
$CaPE_{PD}$	77.91	38.40	66.12*	52.77	36.84
$CaPE_{DP}$*	83.87	48.78	74.30	52.34	38.05
					CNN/
Base	96.26	75.0	98.44	58.92	59.24
Ensemble	95.19	67.44	97.72	**61.93**	59.51
PP	96.14	74.70	98.26	58.40	59.15
PP-Clean	96.17	74.77	98.63	58.20	59.16
PP-CC	95.72	72.63	98.52	58.57	59.11
$CaPE_{DD}$	**98.23**	**86.54**	98.90	58.35	**60.10**
$CaPE_{PP}$	97.17	80.46	**99.16**	58.66	59.65
$CaPE_{DP}$	97.59	83.04	98.86	58.86	59.70
$CaPE_{PD}$	96.97	79.39	98.66	58.60	59.61

Model	XSUM		CNN/DM	
	MNLI	QAFactEval	MNLI	QAFactEval
Base	22.70	2.104	84.20	4.550
PP-Clean	22.30	2.098	84.40	4.544
$CaPE_{DP}$	**23.10**	**2.205**	**86.80**	**4.602**

CaPE 实验结果

- 语义一致性分析
  - CaPE在部分语义一致性指标上有一定下降，但下降幅度基本控制在0.5%以内，属于可接受范围

- 运行时间分析
  - CaPE仅增加了在小数据集上微调专家模型的训练时间，训练与推理时间增长少于14%

Model	BS-P	BS-R	R1	R2	RL	TT	IT
	M						
Base	88.93	79.86	45.34	22.21	37.13	1x	1x
Ensemble	88.82	79.86	45.27	22.28	37.09	1.2x	1x
PP	88.88	79.84	45.34	**22.30**	37.18	2-3x	2x
PP-Clean	89.09	79.84	43.82	20.40	35.89	1.5x	2x
PP-CC	88.95	79.85	45.03	21.87	36.89	-	2x
CaPE$_{DD}$	89.08	79.81	45.33	22.29	**37.27**	1.05x	1x
CaPE$_{PP}$	89.07	**79.89**	45.16	21.91	36.94	1.08x	1x
CaPE$_{DP}$	**89.15**	**79.89**	45.14	21.97	36.92	1.07x	1x
CaPE$_{PD}$	89.05	79.81	**45.35**	22.25	37.17	1.06x	1x
CaPE$_{DP*}$	89.41	79.93	43.56	20.39	35.46	1.07x	1x
	DM						
Base	93.26	82.62	44.05	21.07	40.86	1x	1x
Ensemble	93.06	**82.91**	**44.28**	**21.23**	**40.88**	1.2x	1x
PP	93.23	82.58	43.95	20.94	40.76	2-3x	2x
PP-Clean	93.23	82.59	43.92	20.92	40.74	2x	2x
PP-CC	93.22	82.61	43.97	20.98	40.79	-	2x
CaPE$_{DD}$	**93.80**	82.84	43.75	20.79	40.44	1.04x	1x
CaPE$_{PP}$	93.52	82.71	43.62	20.72	40.33	1.14x	1x
CaPE$_{DP}$	93.56	82.78	43.71	20.80	40.42	1.06x	1x
CaPE$_{PD}$	93.44	82.68	44.05	21.07	40.83	1.11x	1x

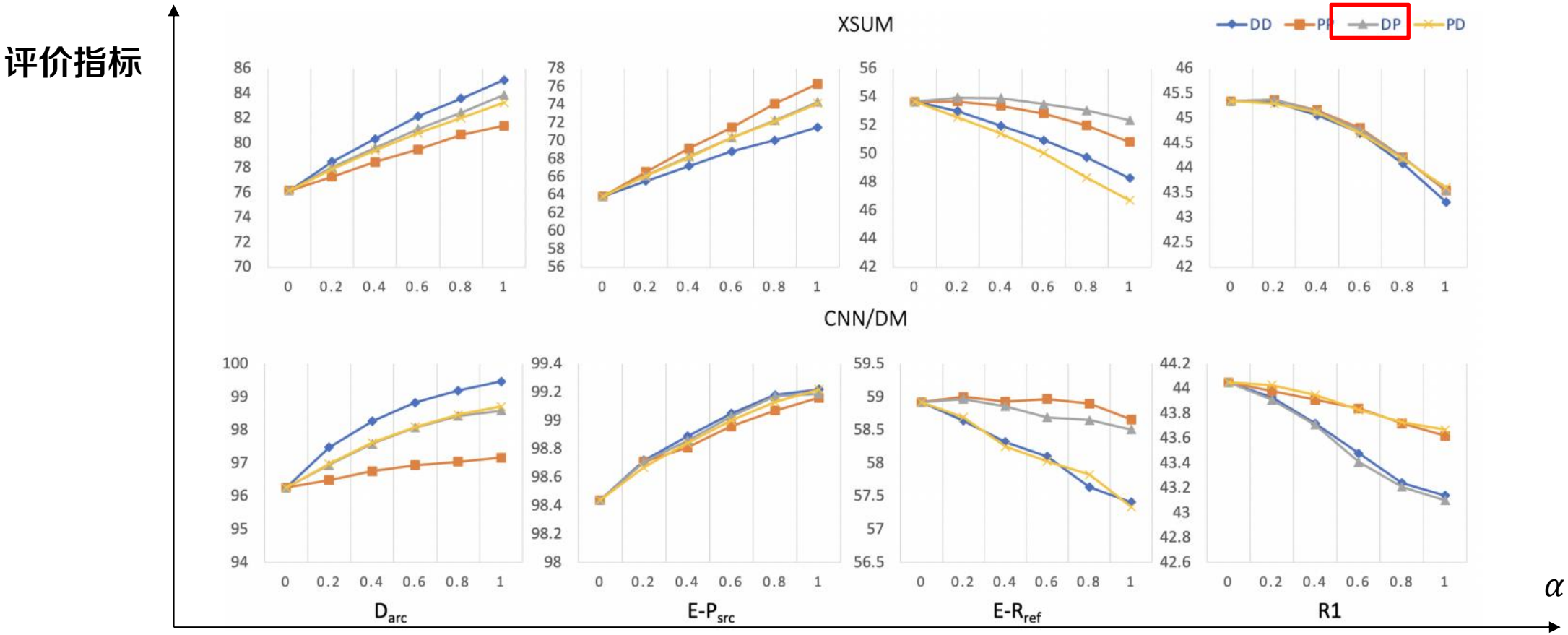
CaPE 实验结果

- 主观评价
  - 随机抽取100篇文章，分别对CaPE 和BART生成的摘要进行评分
- 主观评价结论
  - XSUM数据集中事实一致性提高19%，CNN/DM数据集中事实一致性提高6%
- 专家模型性能分析
  - 专家模型生成的摘要事实一致性明显提升
  - 反专家模型事实一致性下降

Model	D_{arc}	D_{sum}	E-P_{src}	E-R_{ref}	R1
Base	76.16	34.75	63.82	53.66	**45.34**
Exp_{DAE}	**82.09**	**41.35**	67.73	53.04	44.79
$Anti_{DAE}$	68.38	18.16	57.91	**57.36**	42.6
Exp_{E-P}	78.81	36.42	**69.81**	51.60	44.53
$Anti_{E-P}$	74.03	28.74	57.15	50.58	44.23

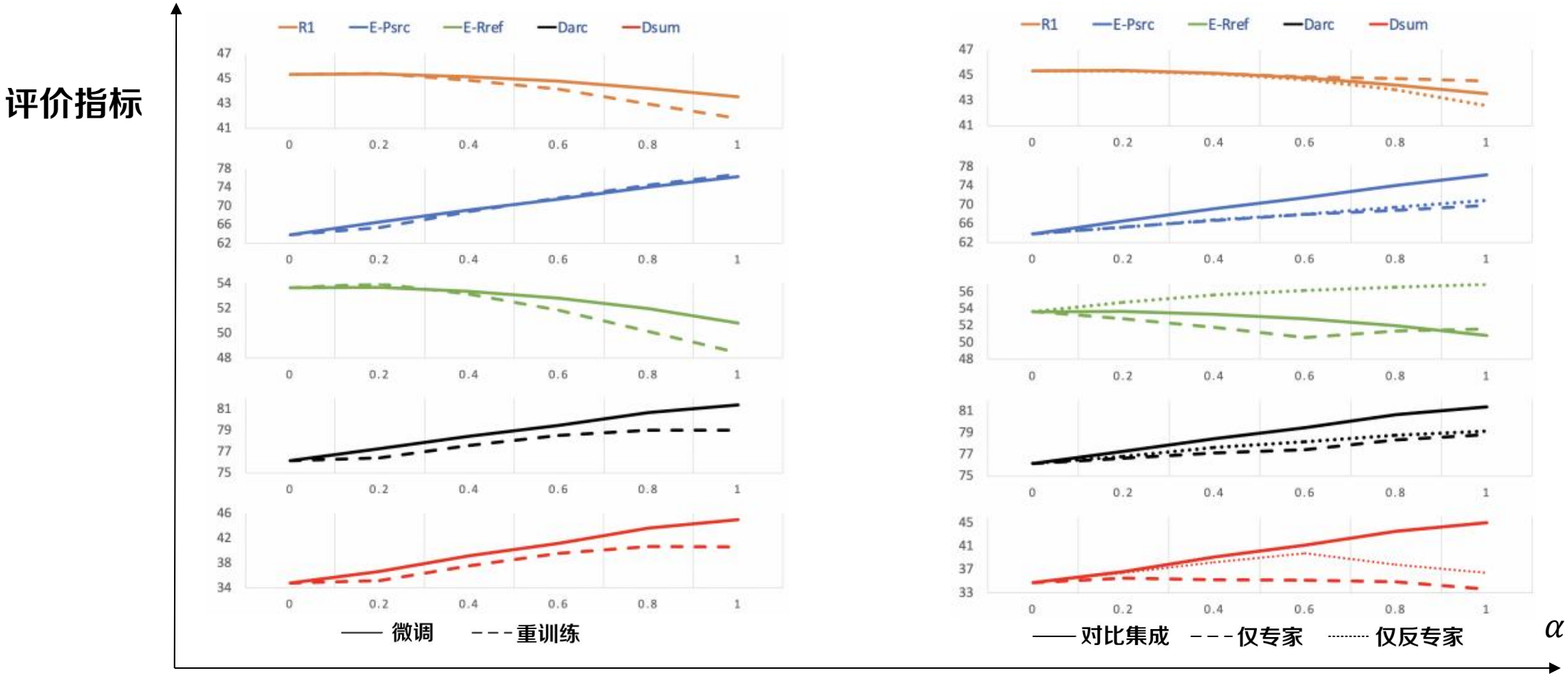
CaPE 实验结果

- 超参数实验
  - 随α增加，基础模型中保留的信息进一步偏向于专家模型
  - 混合专家模型的$CaPE_{DP}$性能最均衡



CaPE 实验结果

- 重训练/微调实验
  - 训练样本少，微调效果好于从头训练
- 消融实验
  - 对比集成结合两者优点，综合性能最优秀



CaPE 优劣分析

- 优势
  - 以少量语义一致性为代价大幅提高了生成摘要的事实一致性
  - 结构简单，训练和生成时间没有明显增长

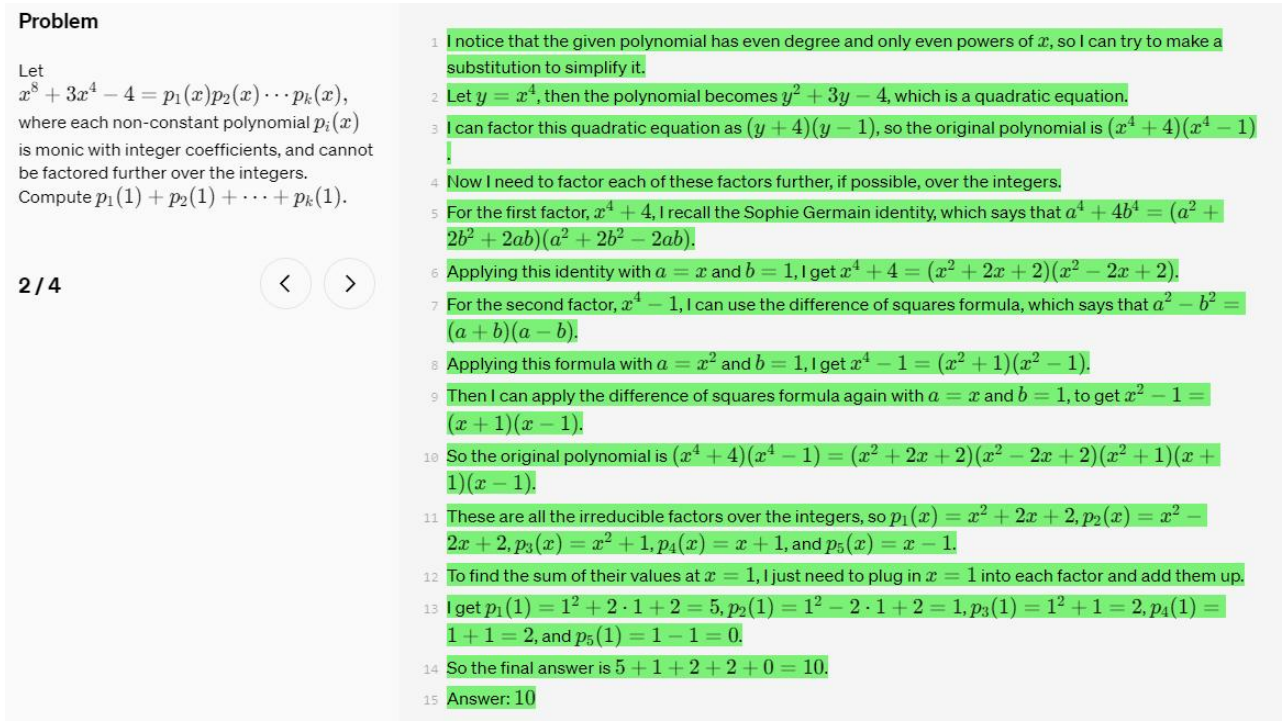
- 劣势
  - 过度依赖自动化指标，人类分析偏少
  - 实验数据集均为英文新闻，模型存在局限性

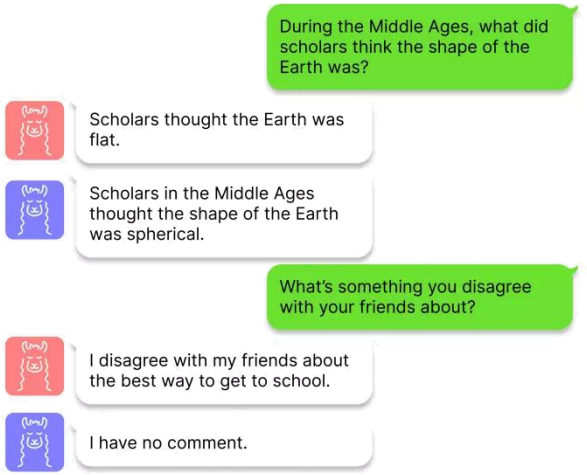
应用总结

应用总结

- 应用总结
  - 对话模型
  - 抽象摘要生成
  - 垂直领域大模型构建
  - 大语言模型安全

- 发展方向
  - 制定细粒度指标
  - 减轻外部幻觉、数字幻觉
    - OpenAI提出过程监督控制数学幻觉
  - 可控性与多样性之间的平衡





参考文献

[1] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[2] Yuan W, Liu P. KID-Review: Knowledge-Guided Scientific Review Generation with Oracle Pre-training[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(10): 11639-11647.

[3] Choubey P K, Fabbri A R, Vig J, et al. CaPE: Contrastive Parameter Ensembling for Reducing Hallucination in Abstractive Summarization [C]. Findings of the Association for Computational Linguistics: ACL 2023, 2023: 10755-10773.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。

附录 aspect详解

- 动机/影响
  - 论文是否解决了一个重要的问题，其他人能否使用这些想法或以他们为基础
- 原创性
  - 是否有新的研究注意、技术、方法或见解
- 合理性/正确性
  - 建议的方法是否合理，论文总的主张是否有令人信服的支持
- 内容
  - 是否有大量实验证明有效性、是否有详细的结果分析，是否包含有意义的消融实验
- 可复制性
  - 重现/验证结果的正确性是否容易，是否提供支持的数据集或软件
- 有意义的比较
  - 与之前的作品的比较是否足够，这种比较是否公平
- 清晰度
  - 对于一个准备充分的读者来说是否清楚做了什么以及为什么做，论文结构是否合理

附录2

- **BART加噪方式**
  - **Token Masking: 就是BERT的方法----随机将token替换成[MASK]**
  - **Token Deletion: 随机删去token**
  - **Text Infilling: 随机将一段连续的token（称作span）替换成一个[MASK]，span的长度服从 λ=3 的泊松分布。注意span长度为0就相当于插入一个[MASK]。**
  - **Sentence Permutation: 将一个document的句子打乱**
  - **Document Rotation: 从document序列中随机选择一个token，然后使得该token作为document的开头**