面向可信自然语言处理的知识概念挖掘与应用方法研究

姓名：刘 烨

学院：人工智能与数据科学学院

专业：数据科学（计算机科学与技术）

导师：陈恩红 教授

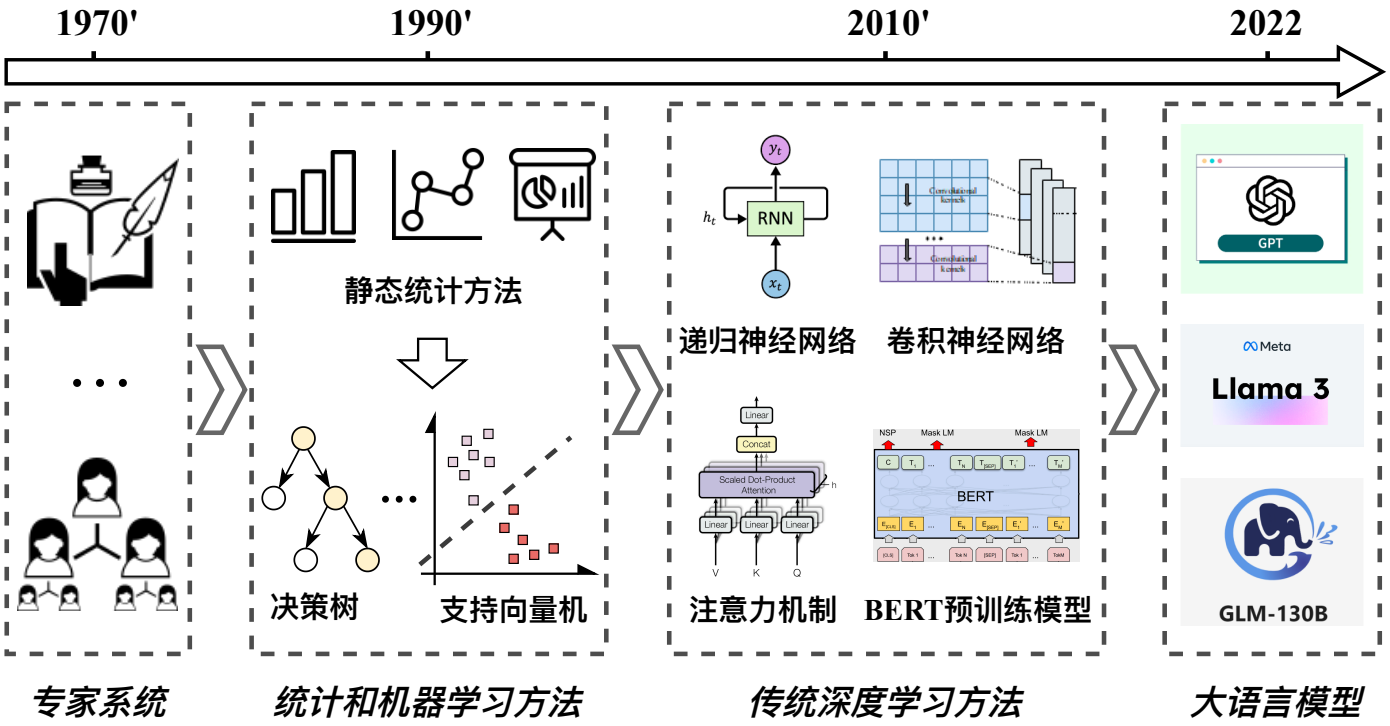
答辩日期：2025 年 3 月 3 日

汇报提纲

1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. 知识概念嵌入增强的层级文本分类方法
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

研究背景

- **自然语言处理技术** 是人工智能领域中的重要分支
  - 旨在利用算法程序自动化处理人们工作生活中的 **文本信息**，将自然语言转化为机器可以 **理解和应用** 的结构化信息，为实现 **更高效**、**更个性化** 的人工智能系统提供技术支持。



研究背景

- **可信自然语言处理（Trustworthy NLP）**
  - 自然语言处理算法的一个重要发展方向
  - 强调 **高准确性**、**高置信度** 和 **可解释性**
- 自2021年，ACL每年均举办主题为 **可信自然语言处理** 的学术研讨会
- 中国科学院 **张钹院士** 在**《加速行业智能化白皮书》**中着重提及了
  - 将 **知识驱动** 和 **数据驱动** 融合，发展 **安全**、**可信**、**可靠** 的AI技术
- 清华大学等发布的**《可信 AI 技术和应用进展白皮书（2023）》**
  - 构建 **融合知识** 的自然语言处理算法对实现可信AI技术有重要意义

Association for Computational Linguistics

研究背景

- 现有 **可信自然语言处理算法** 的相关研究
  - **训练阶段的数据选择方法**
  - 在训练阶段，选择高质量、分布均衡的数据，获得表现稳定的算法模型，增强模型的鲁棒性、准确性
  - **可解释的自然语言处理算法**
  - 对模型的参数进行具体解释：注意力机制可视化；
  - 训练单独的可解释性模块：生成解释性文本，提供预测的理由

研究背景

- 现有 **可信自然语言处理算法** 的相关研究
  - **训练阶段的数据选择方法**
  - 在训练阶段，选择高质量、分布均衡的数据，获得表现稳定的算法模型，增强模型的鲁棒性、准确性

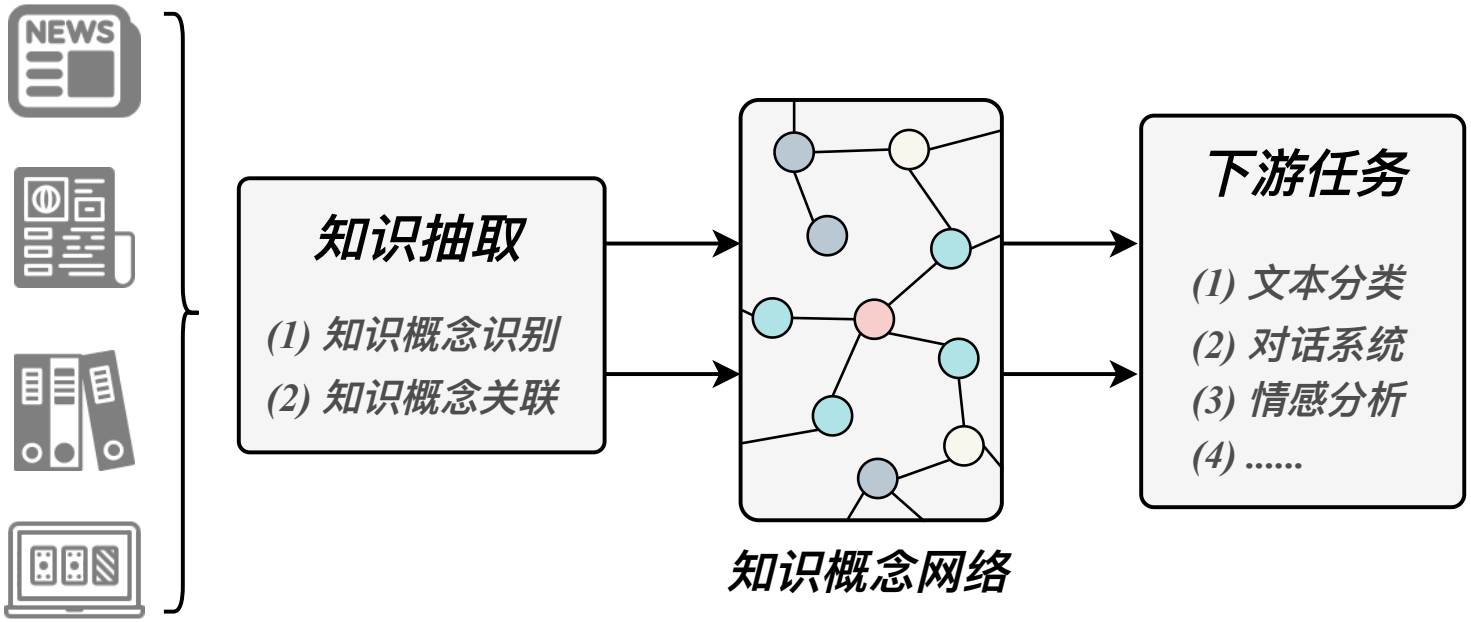
→ **仅解决训练数据分布不均衡导致的问题，受限于已有的训练数据**

  - **可解释的自然语言处理算法**
  - 对模型的参数进行具体解释：注意力机制可视化；
  - 训练单独的可解释性模块：生成解释性文本，提供预测的理由

→ **基于模型本身的逻辑推理能力，模型缺乏特定知识难以发挥作用**

研究背景

- **知识概念挖掘与应用**，基于 **认知科学** 的数据挖掘技术
  - 从大量的文本数据中构建 **知识概念网络**
  - 为下游任务提供相关知识，解决 **领域知识缺乏** 的挑战
  - **显式**、**透明** 的“知识”使用，提高模型的 **可信度** 与 **可解释性**。



研究背景

面向可信自然语言处理的知识概念挖掘与应用方法研究

研究背景	现有的自然语言处理技术存在**领域知识缺乏**、**可信度低**、**可解释性差**等问题，难以构建**可信的自然语言处理算法**

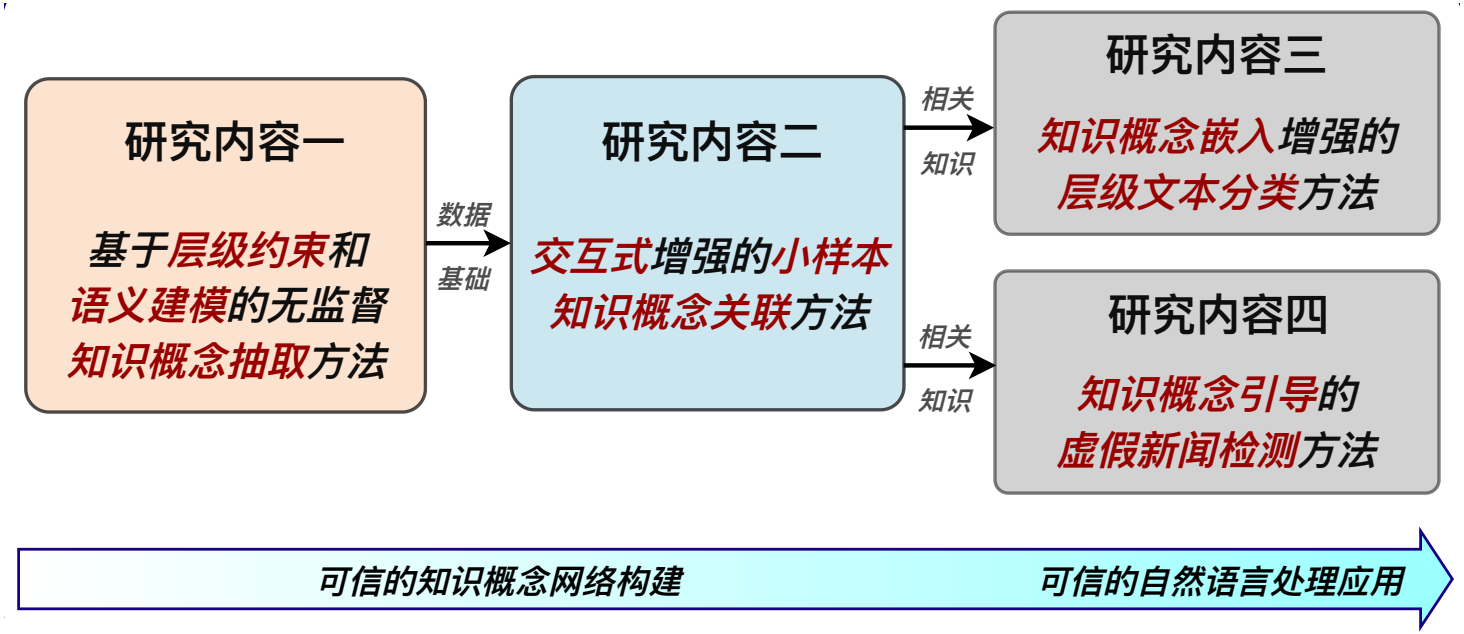
“面向可信自然语言处理的知识概念挖掘与应用方法研究”的**若干挑战**

知识概念语义复杂 解析识别难	知识概念分布分散 关联建模难	知识概念场景多样 下游应用难

- **知识概念语义复杂，解析识别难**
  - 文档中知识概念的各种特征纷繁复杂，难有统一的抽取方法
- **知识概念分布分散，关联建模难**
  - 知识概念间没有直接联系，难以构建结构化的知识概念网络
- **知识概念场景多样，下游应用难**
  - 多种多样的任务场景对知识概念的需求不同，难有统一应用范式

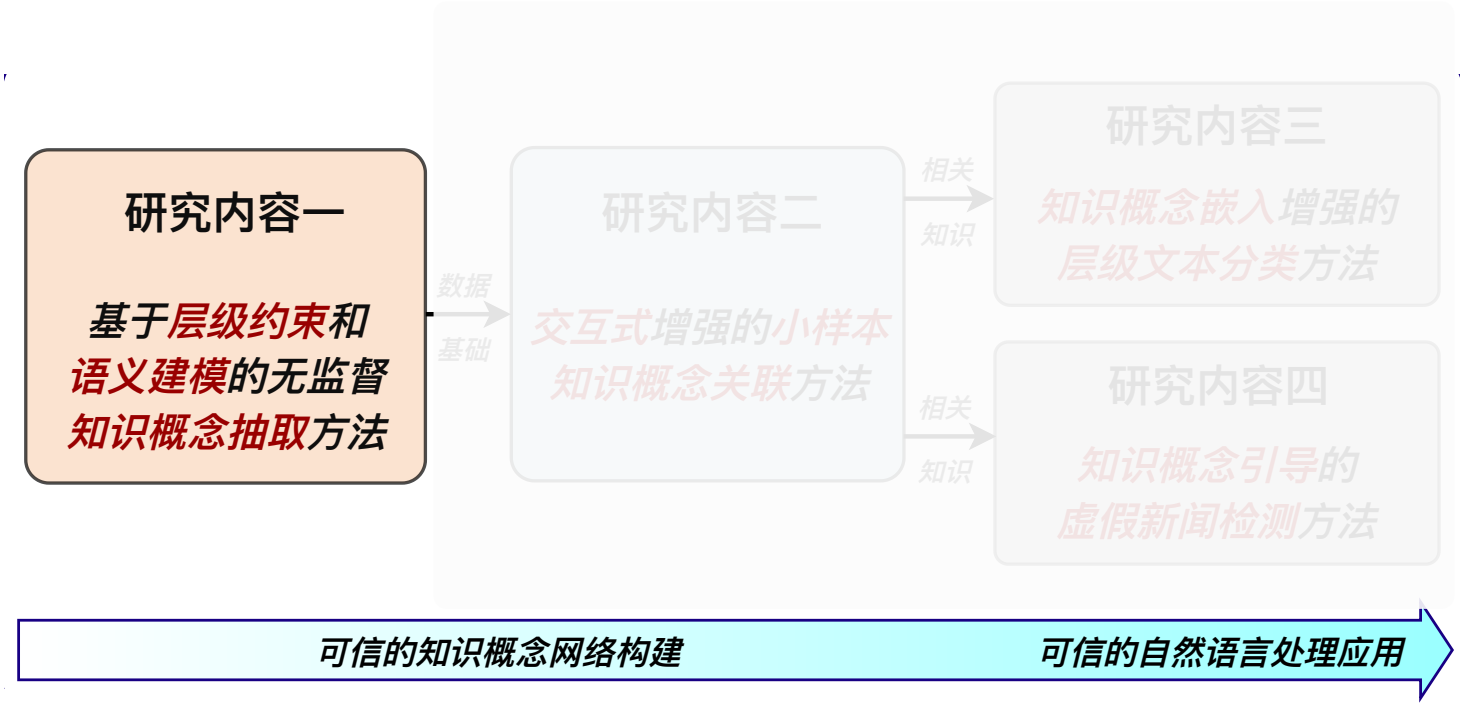
研究背景

- 围绕 **知识概念的抽取**、**关联** 与 **应用** 开展技术研究
  - 形成 **基于层级约束和语义建模的无监督知识概念抽取方法** -- **交互式增强的小样本知识概念关联方法** -- **知识概念嵌入增强的层级文本分类方法** -- **知识概念引导的虚假新闻检测方法** 的研究框架



研究背景

- 围绕 **知识概念的抽取**、**关联** 与 **应用** 开展技术研究
  - 形成 **基于层级约束和语义建模的无监督知识概念抽取方法** -- **交互式增强的小样本知识概念关联方法** -- **知识概念嵌入增强的层级文本分类方法** -- **知识概念引导的虚假新闻检测方法** 的研究框架



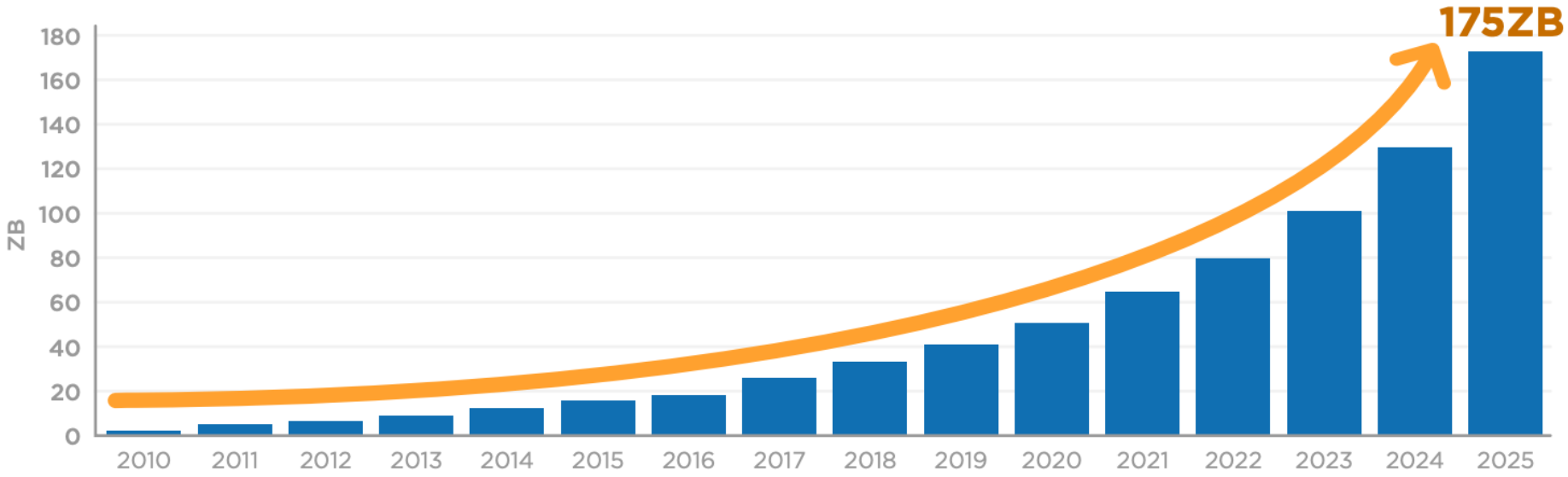
汇报提纲

1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. 知识概念嵌入增强的层级文本分类方法
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

基于层级约束和语义建模的无监督知识概念抽取方法

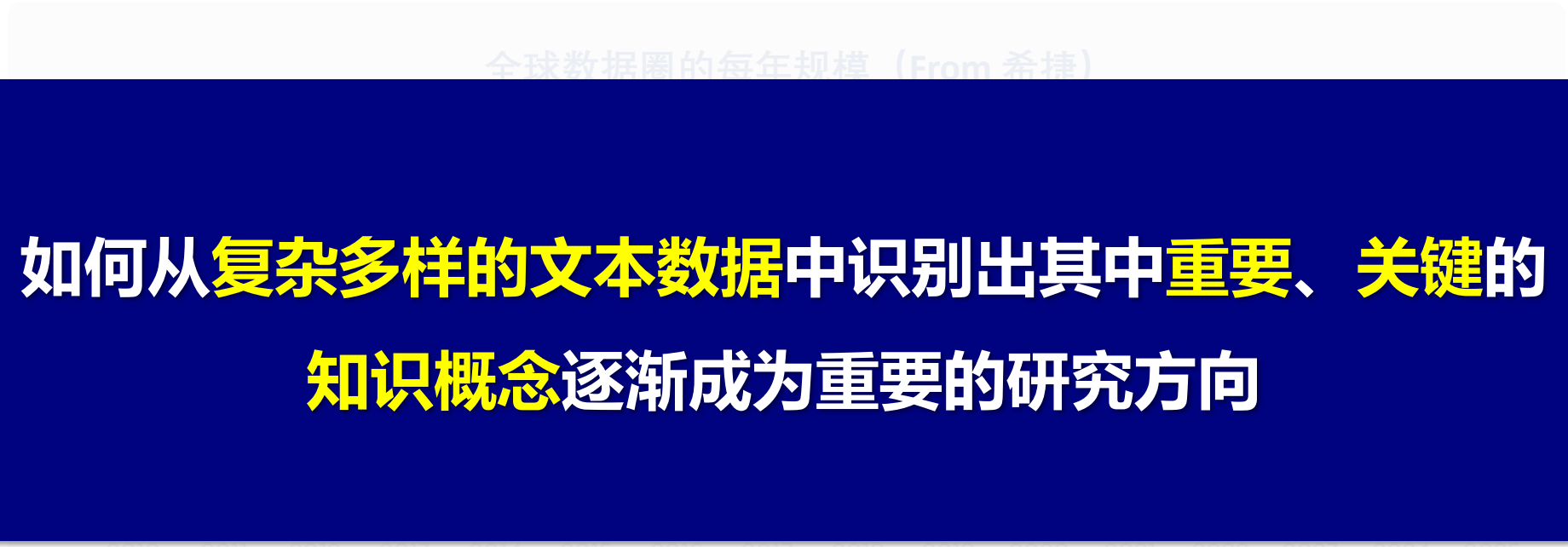
- 当今信息爆炸的时代
  - 全球的数据规模在 2025 年末将达到 **175ZB**（ZB：十万亿亿字节）
  - 每天产生的数据，其中占比高达 **60-73%** 的数据由于各种原因没有被用于有效分析

全球数据圈的每年规模（From 希捷）



基于层级约束和语义建模的无监督知识概念抽取方法

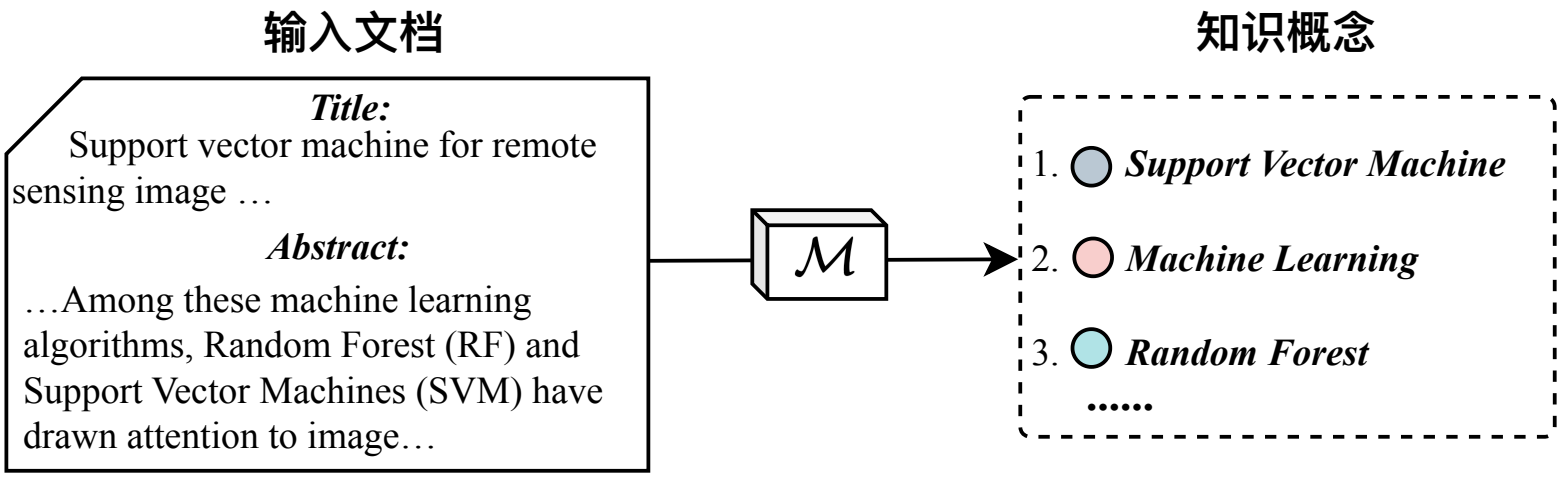
- 当今信息爆炸的时代
  - 全球的数据规模在 2025 年末将达到 **175ZB**（ZB：十万亿亿字节）
  - 每天产生的数据，其中占比高达 **60-73%** 的数据由于各种原因没有被用于有效分析



如何从复杂多样的文本数据中识别出其中重要、关键的知识概念逐渐成为重要的研究方向

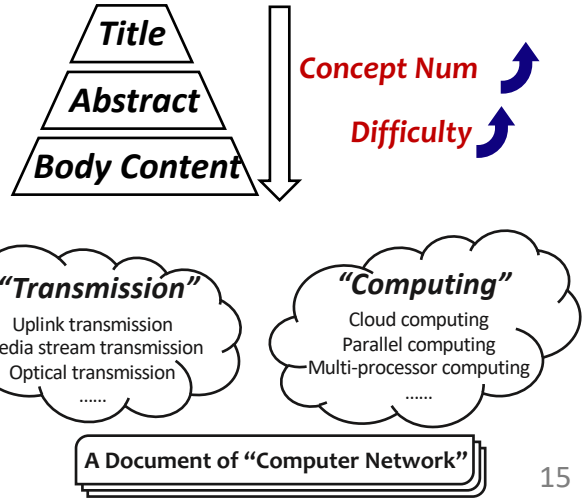
基于层级约束和语义建模的无监督知识概念抽取方法

- 知识概念抽取算法旨在识别出其中的关键知识概念
  - 如：**Support Vector Machine**、**Machine Learning**
  - **知识概念** 通常由原文本中若干个连续的词汇组成，但其不同于一般的短语或命名实体，它们通常具有较强的 **领域相关性**



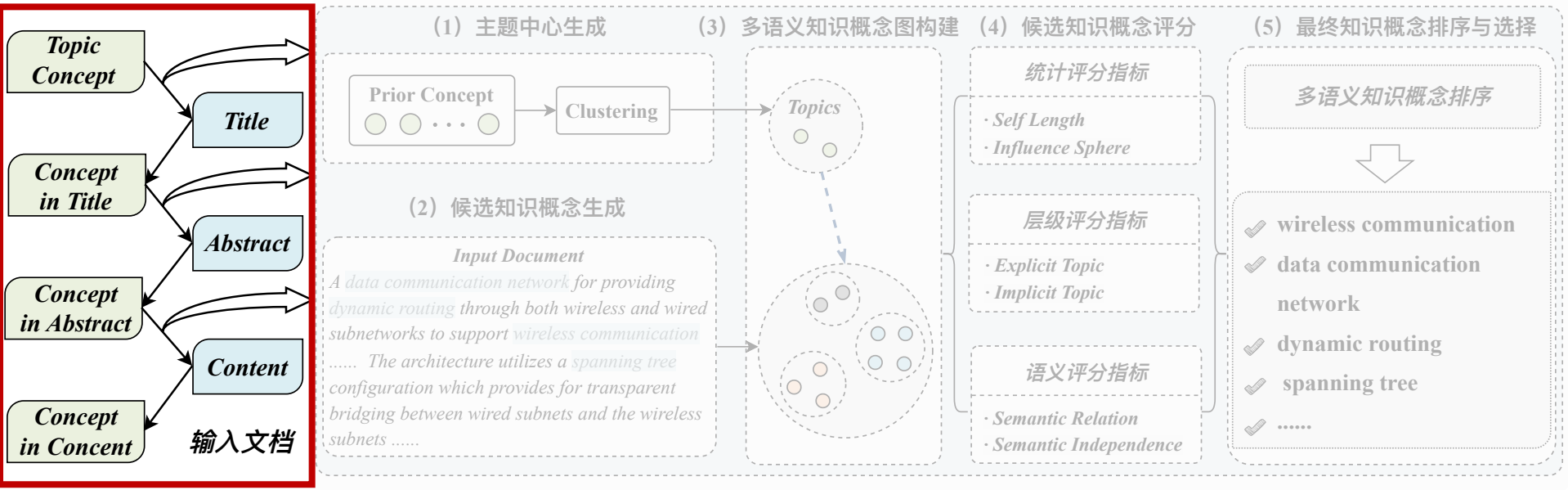
基于层级约束和语义建模的无监督知识概念抽取方法

- 现有的知识概念抽取算法
  - **特征工程方法**：Autophrase (TKDE’ 2018)
    - ✓ 引入远程质量监督机制，借助跨域数据进行特征分类训练
  - **预训练方法**：JMLGC (EMNLP’ 2021)
    - ✓ 使用预训练模型（BERT），挖掘文本中的深层语义特征
- 缺陷
  - 缺乏对文档多层级结构的考虑（Title, Abstract等）
  - 知识概念数量 ⬆ 抽取难度 ⬆
  - 忽略了在文本中知识概念间的复杂语义关联，尤其在长文本中



基于层级约束和语义建模的无监督知识概念抽取方法

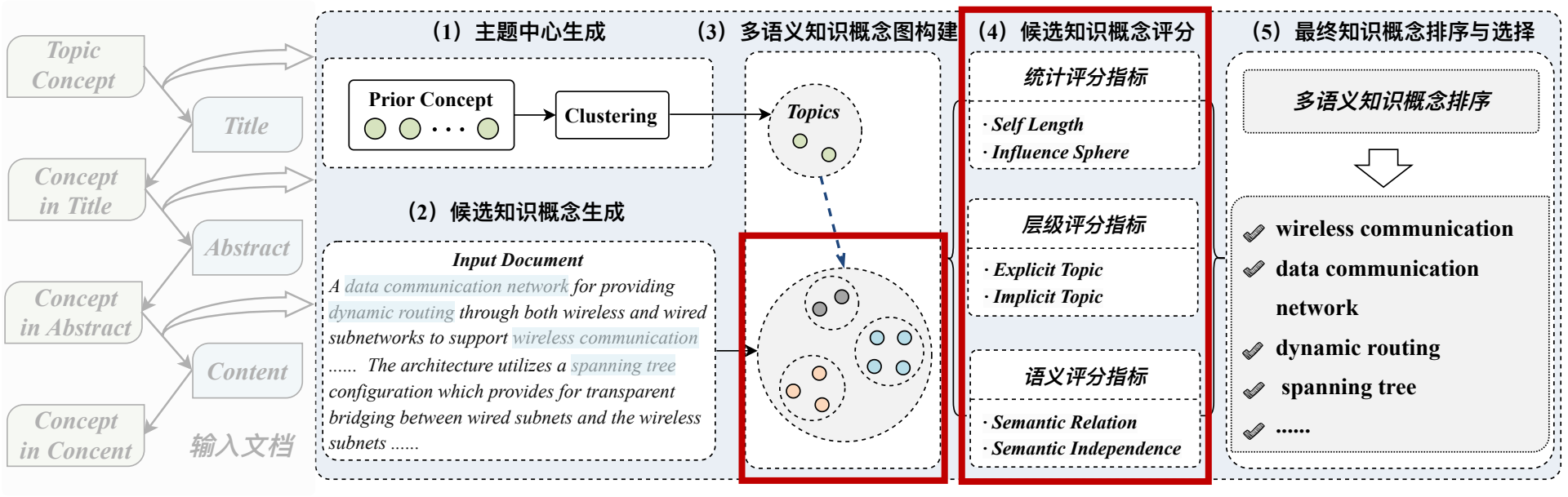
- **层级递进式** 的抽取架构
  - 按照层级 **从上到下** 进行知识概念的抽取，并且上一层级的抽取结果会作为 **指导信息辅助下一层级** 的抽取



层级递进的抽取架构

基于层级约束和语义建模的无监督知识概念抽取方法

- 每一个层级内部，遵循 **先生成候选知识概念**、**再进行筛选** 的过程
  - **多语义知识概念图** – 建模文本中不同知识概念间的关系
  - **多角度的知识概念评分指标** – 细致考量知识概念的多种特征



基于层级约束和语义建模的无监督知识概念抽取方法

- 每一个层级内部，遵循 **先生成候选知识概念**、**再进行筛选** 的过程
  - **多语义知识概念排序算法** – 在多语义知识概念图上，从子图和全局两个角度进行影响力的传播

$$R(v_i) = (1-d) \cdot I(v_i) + d \cdot I(v_i) \cdot ((1-\gamma) \cdot R(v_i)_{local} + \gamma \cdot R(v_i)_{global})$$

Algorithm 1 多语义知识概念排序算法

Input: 多语义知识概念图 $G = (V, S, E, W)$；节点的归一化得分 $I(v_i)$；阻尼因子 d；调和因子 γ

Output: 排序后的知识概念列表 P_{list}

1: 初始化列表 R_{list} 为均匀分布
2: **while** 未收敛 **do**
3: 计算每个子图 $s_i \in S$ 的排名值 $R(s_i)$，参见公式 (3.12)。
4: 从局部和全局视角更新 R_{list}，参见公式 (3.13 - 3.15)。
5: **end while**
6: 根据 R_{list} 对所有候选知识概念进行排序，生成排序后的知识概念列表 P_{list}
7: **return** P_{list}

基于层级约束和语义建模的无监督知识概念抽取方法

实验数据集

- 美国专利商标局（USPTO）公开文档数据
- Mechanical Engineering（机械工程）
  - 机械工程、照明、加热、武器、爆破装置或泵 相关的专利文档
- Electricity（电学）
  - 涉及电学相关的文档

数据集	文档数量	Title 平均句子数量	Abstract 平均句子数量	Content 平均句子数量
Mechanical Engineering	11,186	1.00	3.85	13.58
Electricity	84,069	1.00	3.89	16.58

基于层级约束和语义建模的无监督知识概念抽取方法

实验结果

- 提出的 TechPat 模型在所有指标上 **均优于基线模型**
- DBpedia 在 Precision 上表现出色，但由于完全依赖外部数据库且只能提取 **极少量的知识概念**，在 Recall 和 F1-score 上 的表现较差

Method	Mechanical Engineering			Electricity		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ECON	26.70	10.43	14.01	23.76	8.19	11.35
DBpedia	**43.13**	11.49	16.80	35.08	10.29	14.99
Autophrase	28.18	26.83	25.47	27.49	31.83	27.27
NE-rank	20.01	31.05	22.81	21.53	33.23	24.11
Rake	16.17	26.89	18.78	14.03	24.53	16.53
Spacy	32.42	48.83	36.41	32.37	49.27	36.20
MultipartiteRank	37.80	51.21	40.66	36.37	49.15	38.84
JMLGC	34.86	48.58	37.92	37.67	50.05	39.92
TechPat	39.83	**55.32**	**43.10**	**38.98**	**55.10**	**41.89**

基于层级约束和语义建模的无监督知识概念抽取方法

实验结果

- 分析在 **不同层级** 上的表现：Title、Abstract、Content
- 在 **长文本层级**（Content）上的领先幅度，要远显著于 **短文本层级**（Title），更证明了所提出方法在知识概念识别任务上的强大能力

Method	Title			Abstract			Content		
	P	R	F1	P	R	F1	P	R	F1
ECON	15.00	9.00	10.90	18.58	7.52	9.49	18.54	7.88	10.42
DBpedia	14.50	9.67	11.17	34.19	12.95	17.09	33.42	10.46	14.80
Autophrase	8.00	4.50	5.50	24.34	13.85	15.98	23.76	28.69	23.68
NE-rank	34.00	38.17	35.17	22.53	36.76	25.05	14.08	19.73	14.73
Rake	59.00	59.00	57.47	21.89	35.72	25.26	5.10	7.01	5.08
Spacy	61.50	58.50	58.43	37.29	53.29	40.28	22.82	28.75	22.33
MultipartiteRank	58.50	53.67	54.83	39.06	51.93	41.36	33.63	44.16	33.58
JMLGC	**63.50**	59.17	59.50	41.27	55.08	43.48	28.48	36.19	27.72
TechPat	61.00	**64.67**	**60.90**	**43.01**	**61.71**	**46.10**	**34.28**	**45.32**	**34.14**

基于层级约束和语义建模的无监督知识概念抽取方法

本章小结

- 针对知识概念抽取任务进行了充分的数据分析
- 提出了 **基于层级约束和语义建模的无监督知识概念抽取方法**
- 引领了 **多层级的抽取范式**，启发了后续诸多抽取模型[1-5]

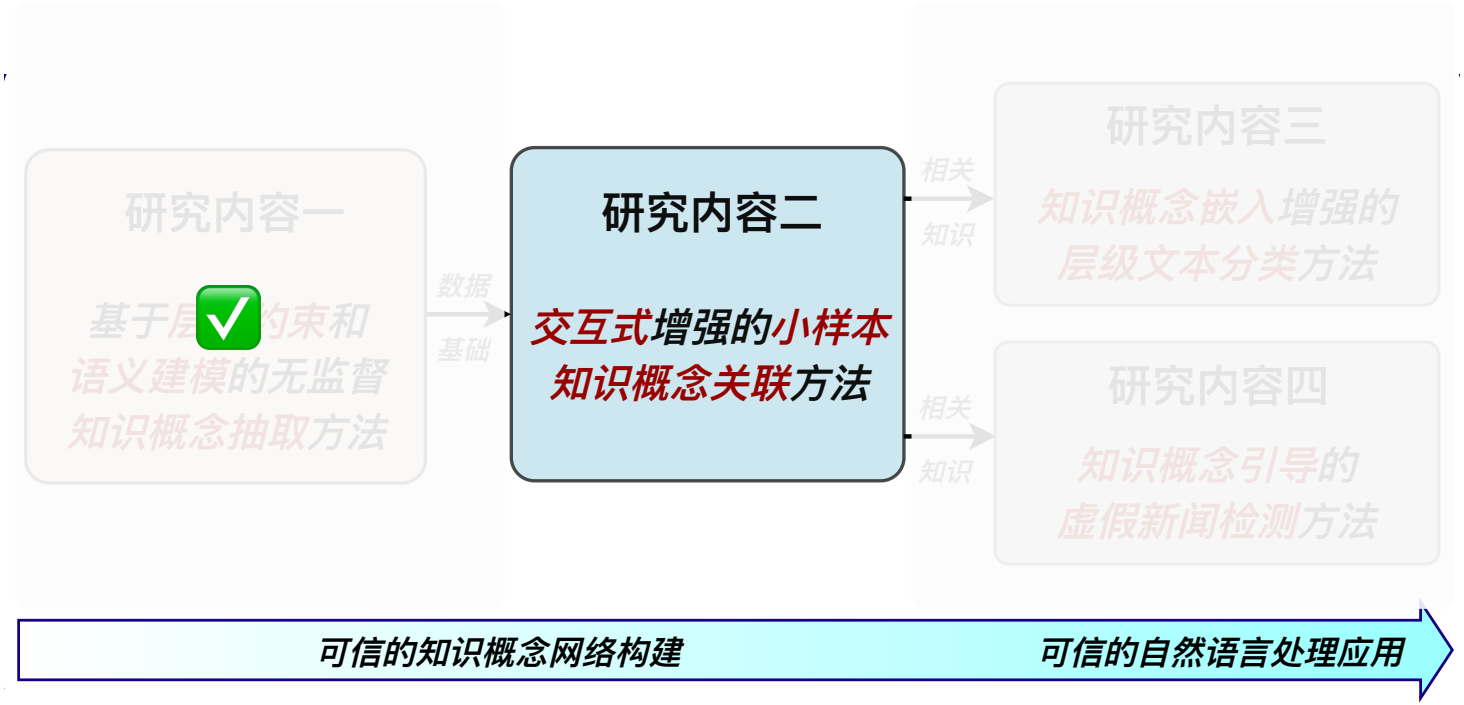
[1] Zhou P, Jiang X, Zhao S. Unsupervised technical phrase extraction by incorporating structure and position information[J]. Expert Systems with Applications, 2024.
[2] Miao R, Chen X, Hu L, et al. PatSTEG: Modeling Formation Dynamics of Patent Citation Networks via The Semantic-Topological Evolutionary Graph[C]//2023 IEEE International Conference on Data Mining (ICDM). IEEE, 2023: 1229-1234.
[3] Mao R, He K, Zhang X, et al. A survey on semantic processing techniques[J]. Information Fusion, 2024, 101: 101988.
[4] Marques T D, Gonçalves A L. UMA REVISÃO INTEGRATIVA PARA SISTEMAS DE BUSCA POR PATENTES SIMILARES UTILIZANDO IA: AVANÇOS, DESAFIOS E APLICAÇÕES[C]//Anais do Congresso Internacional de Conhecimento e Inovação–ciki. 2023.
[5] Gao W, Wang H, Liu Q, et al. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis[C]//Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval. 2023: 983-992.

相关工作以第一作者身份发表于 CCF B 类会议

ICDM2020 和 CCF B 类期刊 ACM TKDD2023 上

基于层级约束和语义建模的无监督知识概念抽取方法

- 围绕 **知识概念的抽取**、**关联** 与 **应用** 开展技术研究
  - 形成 **基于层级约束和语义建模的无监督知识概念抽取方法** -- **交互式增强的小样本知识概念关联方法** -- **知识概念嵌入增强的层级文本分类方法** -- **知识概念引导的虚假新闻检测方法** 的研究框架

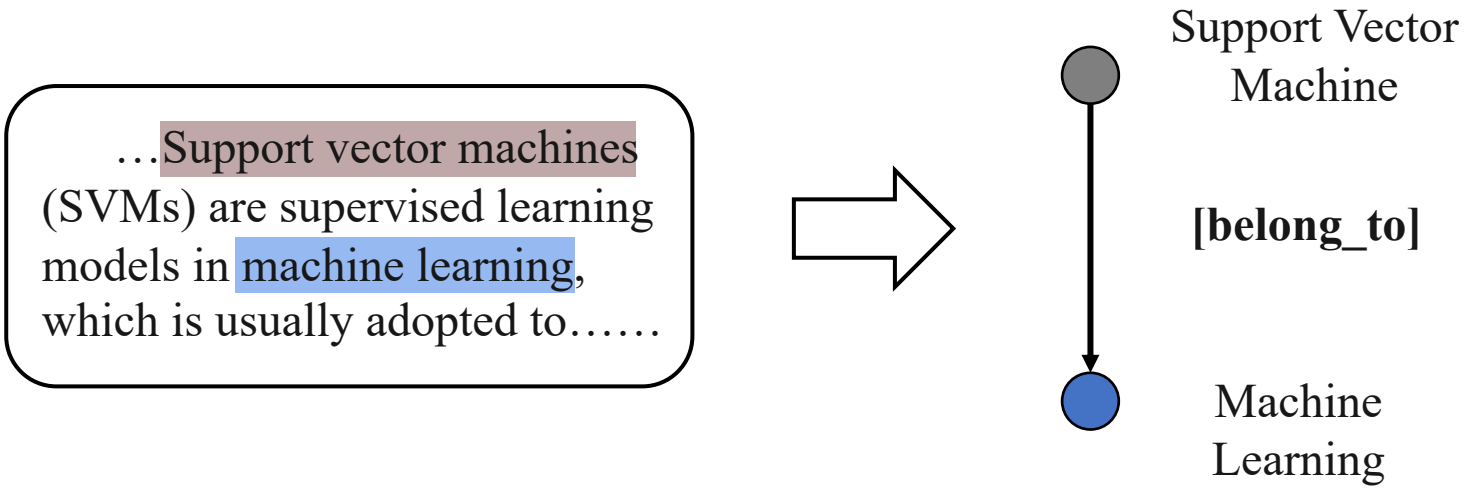


汇报提纲

1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. 知识概念嵌入增强的层级文本分类方法
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

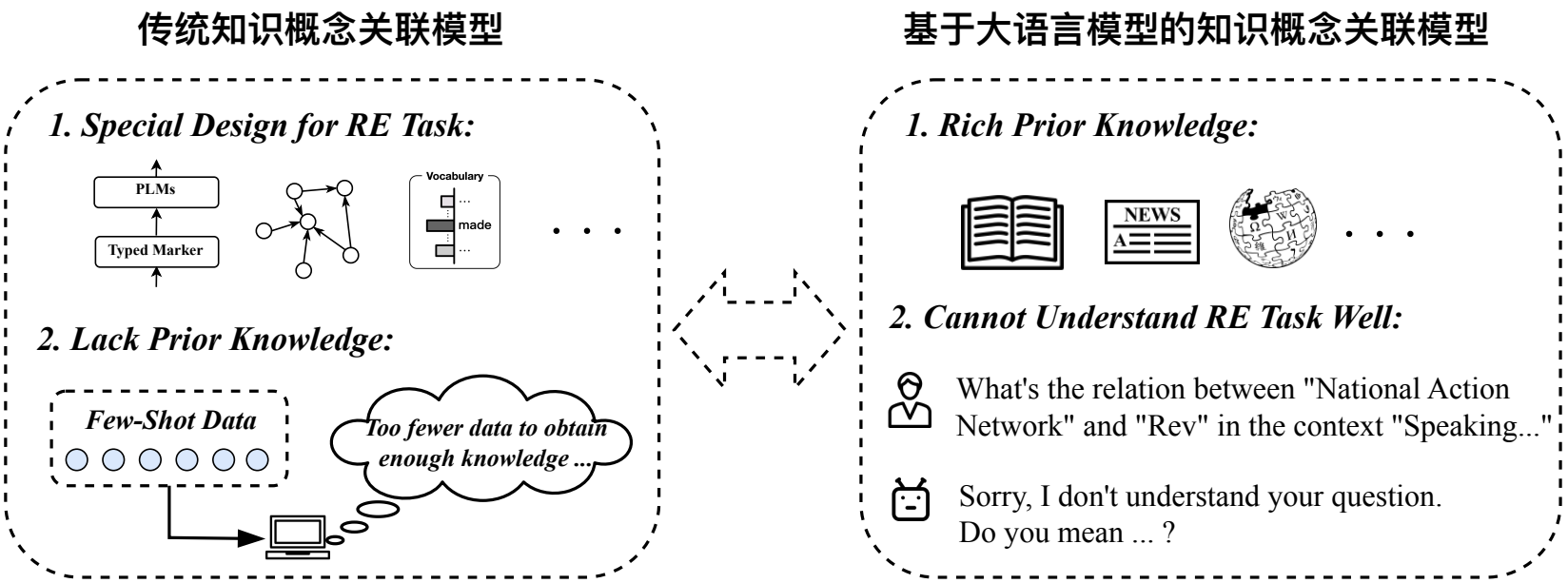
交互式增强的小样本知识概念关联方法

- **知识概念关联**
  - 给定从文本中抽取得到的 **知识概念对 (c_1, c_2)**，预测两个知识概念之间的 **关联关系：$r \in R$**，其中 R 是预先定义好的关系集合
  - 例子：



交互式增强的小样本知识概念关联方法

- 相关工作
  - **传统知识概念关联模型**：KnowPrompt（WWW’ 2022）
    - ✓ 为知识概念关联任务“**量身定制**”，但解决问题的 **先验知识** 不足
  - **基于大语言模型的方法**：Unleash (ACL’ 2023 Workshop)
    - ✓ 拥有大量的 **先验知识**，但对于 **特定任务** 的理解和推理能力不足



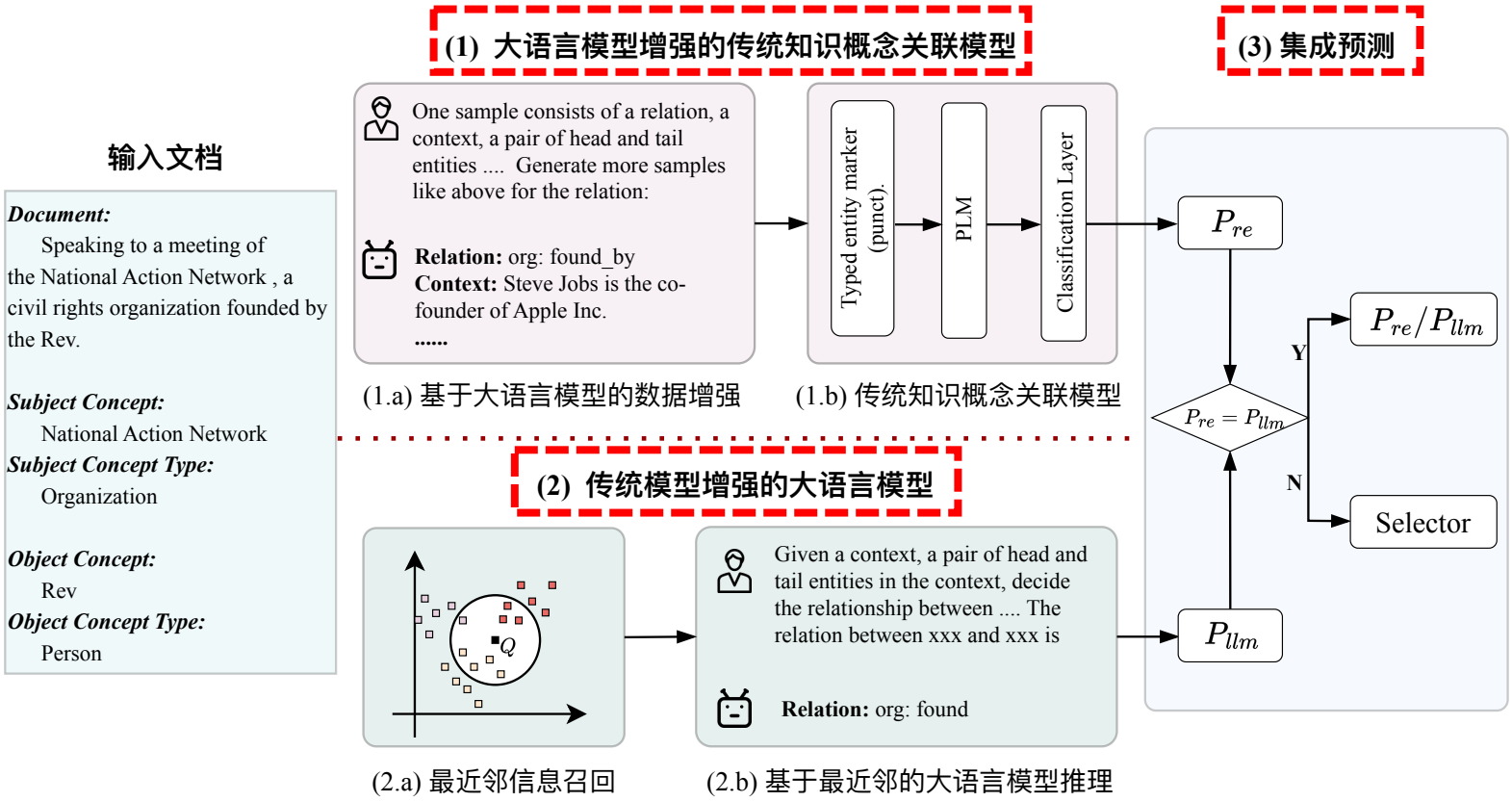
交互式增强的小样本知识概念关联方法

- 相关工作
  - **传统知识概念关联模型**：KnowPrompt（WWW’ 2022）
    - ✓ 为知识概念关联任务“**量身定制**”，但解决问题的 **先验知识** 不足
  - **基于大语言模型的方法**：Unleash (ACL’ 2023 Workshop)
    - ✓ 拥有大量的 **先验知识**，但对于 **特定任务** 的理解和推理能力不足

传统知识概念关联模型和基于大语言模型的方法的优劣势相互对立，如何设计综合方法优势互补？

交互式增强的小样本知识概念关联方法

- 提出了一种将 **传统知识概念关联方法** 和 **大语言模型** 结合的新思路
  - 将大语言模型的 **先验知识** 注入传统知识概念关联模型
  - 将传统知识概念关联模型对 **此任务的理解** 传递给大语言模型



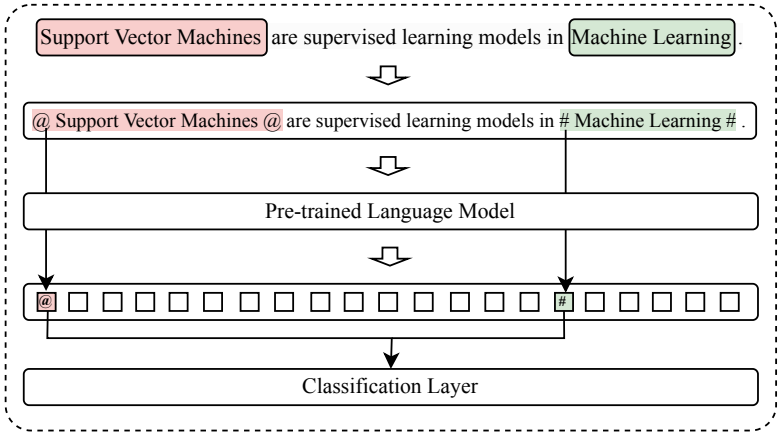
交互式增强的小样本知识概念关联方法

- （1）**大语言模型增强的传统知识概念关联模型**
  - 将大语言模型的 **先验知识** 注入传统知识概念关联模型
  - （1.a）基于大语言模型的数据增强
    - ✓ 引导大语言模型创造更多的 **伪知识概念关联样本**

One sample in relation extraction datasets consists of a relation, a context, a pair of head and tail entities in the context and their entity types. The head entity has the relation with the tail entity and entities are pre-categorized as the following types: [ENTITY TYPE List].
Here are some samples for relation 'RELATION':
Relation: RELATION. Context: TEXT. Head Type: HEAD TYPE. Head Entity: HEAD ENTITY. Tail Type: TAIL TYPE. Tail Entity: TAIL ENTITY. × N
Generate more samples like above for the relation 'RELATION'. _____

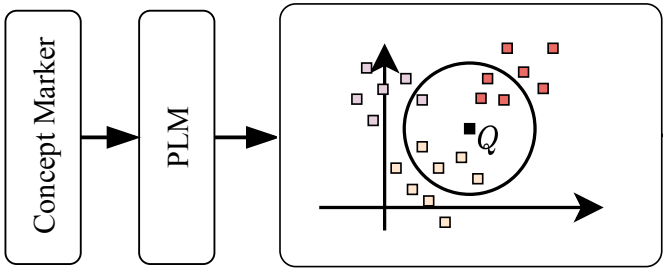
Relation: RELATION.
Context: TEXT.
Head Type: HEAD TYPE.
Head Entity: HEAD ENTITY.
Tail Type: TAIL TYPE.
Tail Entity: TAIL ENTITY
...

  - （1.b）传统知识概念关联模型 → P_{re}



交互式增强的小样本知识概念关联方法

- (2) **传统模型增强的大语言模型**
  - 将传统知识概念关联模型对 **此任务的理解** 传递给大语言模型
  - (2.a) 最近邻信息召回：
    - ✓ 利用 k 最近邻检索方法从训练集中检索更有价值的样本



  - (2.a) 基于最近邻的大语言模型推理 → P_{llm}

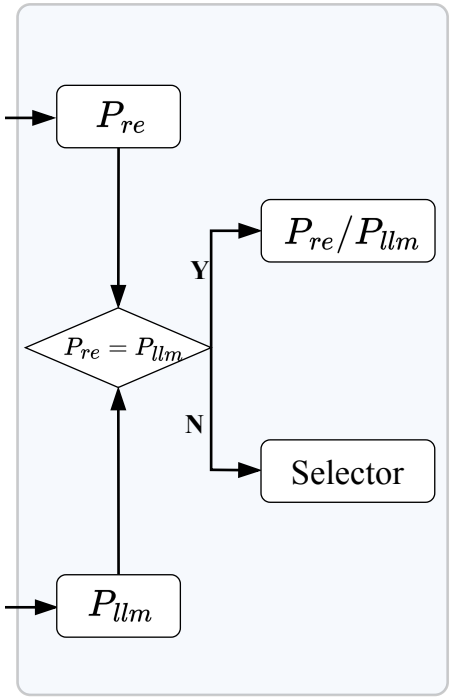
$$P\left(y_{test} \mid P_{inference} \uplus \mathcal{N} \uplus x_{test}\right)$$

Given a context, a pair of head and tail entities in the context, decide the relationship between the head and tail entities from candidate relations: [RELATION List].
Context: TEXT. The relation between (HEAD TYPE) 'HEAD ENTITY' and (TAIL TYPE) 'TAIL ENTITY' in the context is RELATION. × N
Context: TEXT. The relation between (HEAD TYPE) 'HEAD ENTITY' and (TAIL TYPE) 'TAIL ENTITY' in the context is _____

Demonstrations → LLM → Text Completion → RELATION

交互式增强的小样本知识概念关联方法

- (3) **集成预测模块**
  - 两个预测结果相同： $P_{re} = P_{llm}$
    - ✓ 直接返回预测结果
  - 两个预测结果不同： $P_{re} != P_{llm}$
    - ✓ 传统模型和大语言模型存在矛盾
    - ✓ 从训练数据集合中召回 $2m$ 个样本
    - ✓ 得到了最后的预测结果 → P_f



$$P\left(y_{test} \mid P_{inference} \uplus \mathcal{N} \uplus x_{test}\right)$$

Given a context, a pair of head and tail entities in the context, decide the relationship between the head and tail entities from candidate relations: [RELATION List].
Context: TEXT. The relation between (HEAD TYPE) 'HEAD ENTITY' and (TAIL TYPE) 'TAIL ENTITY' in the context is RELATION.
Context: TEXT. The relation between (HEAD TYPE) 'HEAD ENTITY' and (TAIL TYPE) 'TAIL ENTITY' in the context is _____

× 2m Demonstrations

LLM → Text Completion → RELATION

交互式增强的小样本知识概念关联方法

实验数据集

- TACKBP4 挑战赛 -- **知识概念关联数据集**
  - **TACRED**：大规模句子级知识概念关联数据集
  - **TACREV**：修正了原始 TACRED 的开发集和测试集中的错误
  - **Re-TACRED**：重构了 TACRED 的训练集、开发集和测试集
  - **小样本设定**：每种关系随机抽取 K 个实例（K-shot）用于训练和验证阶段，测试集保持完整

Dataset	#Train	#Dev	#Test	#Rel
TACRED	8/16/32	8/16/32	15,509	42
TACREV	8/16/32	8/16/32	15,509	42
Re-TACRED	8/16/32	8/16/32	13,418	40

交互式增强的小样本知识概念关联方法

实验结果

- 提出的 DSARE 模型在所有指标、设定上 **均优于基线模型**
- DSARE 在 **高噪数据**（TACRED、TACREV）上的领先幅度，要显著优于 **低噪数据**（Re-TACREV）

Methods	TACRED			TACREV			Re-TACRED		
	K=8	K=16	K=32	K=8	K=16	K=32	K=8	K=16	K=32
① TYP Marker	29.02	31.35	31.86	26.28	29.24	31.55	51.32	55.60	57.82
② PTR	28.34	29.39	30.45	28.63	29.75	30.79	47.80	53.83	60.99
③ KnowPrompt	30.30	33.53	34.42	30.47	33.54	33.86	56.74	61.90	65.92
④ GenPT	35.45	35.58	35.61	33.81	33.93	36.72	57.03	57.66	65.25
⑤ GPT-3.5		29.72			29.98			39.06	
⑥ LLama-2		22.68			21.96			34.31	
⑦ Zephyr		37.10			38.83			35.81	
⑧ Unleash	32.24	33.81	34.76	32.70	34.53	35.28	58.29	64.37	66.03
DSARE (ours)	**43.84**	**45.40**	**45.94**	**44.69**	**46.61**	**46.94**	**60.04**	**66.83**	**67.13**

交互式增强的小样本知识概念关联方法

实验结果

- 提出的 DSARE 模型在所有指标、设定上 **均优于基线模型**
- DSARE 在 **高噪数据**（TACRED、TACREV）上的领先幅度，要显著优于 **低噪数据**（Re-TACREV）

Noisy Data *Clean Data*

Methods	TACRED			TACREV			Re-TACRED		
	K=8	K=16	K=32	K=8	K=16	K=32	K=8	K=16	K=32
① TYP Marker	29.02	31.35	31.86	26.28	29.24	31.55	51.32	55.60	57.82
② PTR	28.34	29.39	30.45	28.63	29.75	30.79	47.80	53.83	60.99
③ KnowPrompt	30.30	33.53	34.42	30.47	33.54	33.86	56.74	61.90	65.92
④ GenPT	35.45	35.58	35.61	33.81	33.93	36.72	57.03	57.66	65.25
⑤ GPT-3.5		29.72			29.98			39.06	
⑥ LLama-2		22.68			21.96			34.31	
⑦ Zephyr		37.10			38.83			35.81	
⑧ Unleash	32.24	33.81	34.76	32.70	34.53	35.28	58.29	64.37	66.03
DSARE (ours)	**43.84**	**45.40**	**45.94**	**44.69**	**46.61**	**46.94**	**60.04**	**66.83**	**67.13**

交互式增强的小样本知识概念关联方法

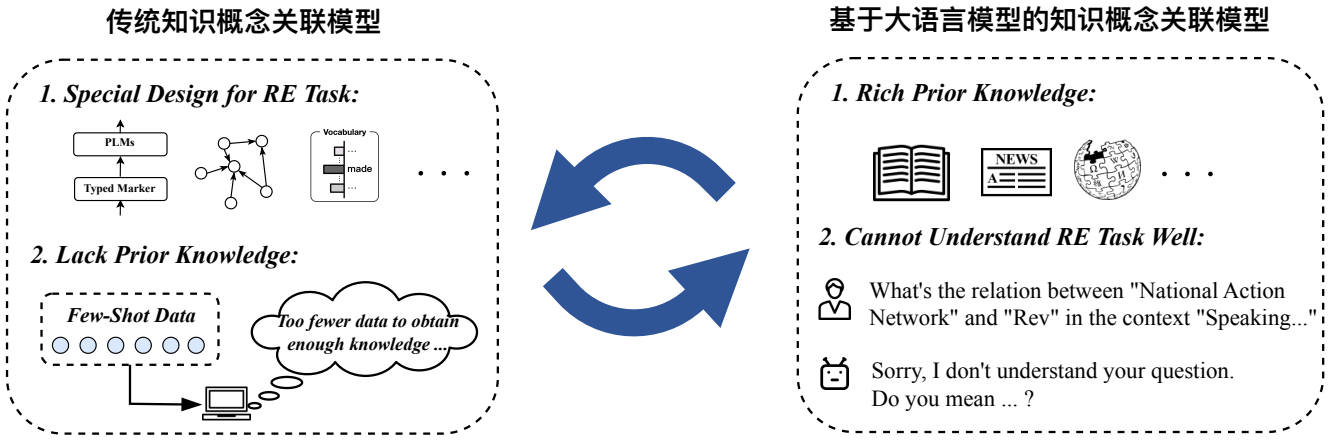
案例分析

- (a) 中，两个消融变体均做出了正确的预测
- (b) 和 (c) 中，借助集成预测模块，DSARE 模型最终得出了正确的预测。

交互式增强的小样本知识概念关联方法

本章小结

- **首次提出** 融合 **传统知识概念关联方法** 和 **大语言模型** 优势
- 提出了 **交互式增强的小样本知识概念关联方法**



相关工作以第一作者身份发表于 CCF B 类会议

DASFAA2024 上

交互式增强的小样本知识概念关联方法

- 围绕 **知识概念的抽取**、**关联** 与 **应用** 开展技术研究
  - 形成 **基于层级约束和语义建模的无监督知识概念抽取方法** -- **交互式增强的小样本知识概念关联方法** -- **知识概念嵌入增强的层级文本分类方法** -- **知识概念引导的虚假新闻检测方法** 的研究框架

汇报提纲

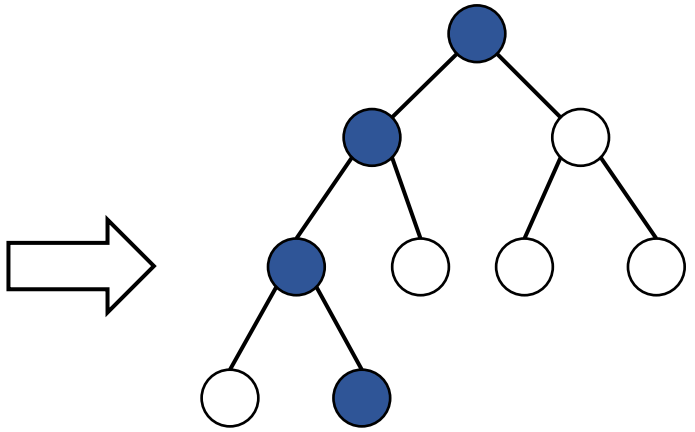
1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. **知识概念嵌入增强的层级文本分类方法**
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

知识概念嵌入增强的层级文本分类方法

- **层级文本分类**（Hierarchical Text Classification, HTC），是一种特殊类型的 **多标签文本分类问题**。在这一任务中，文档会被标注多个类别，而这些类别通常以 **树** 或 **无环图** 的形式构成

- Example：

It is as vast as the USA and so arid that most bacteria cannot survive there. The author came to the Sahara to see it as its inhabitants do, riding its public transport, from Algiers to Dakar



Taxonomic Hierarchy

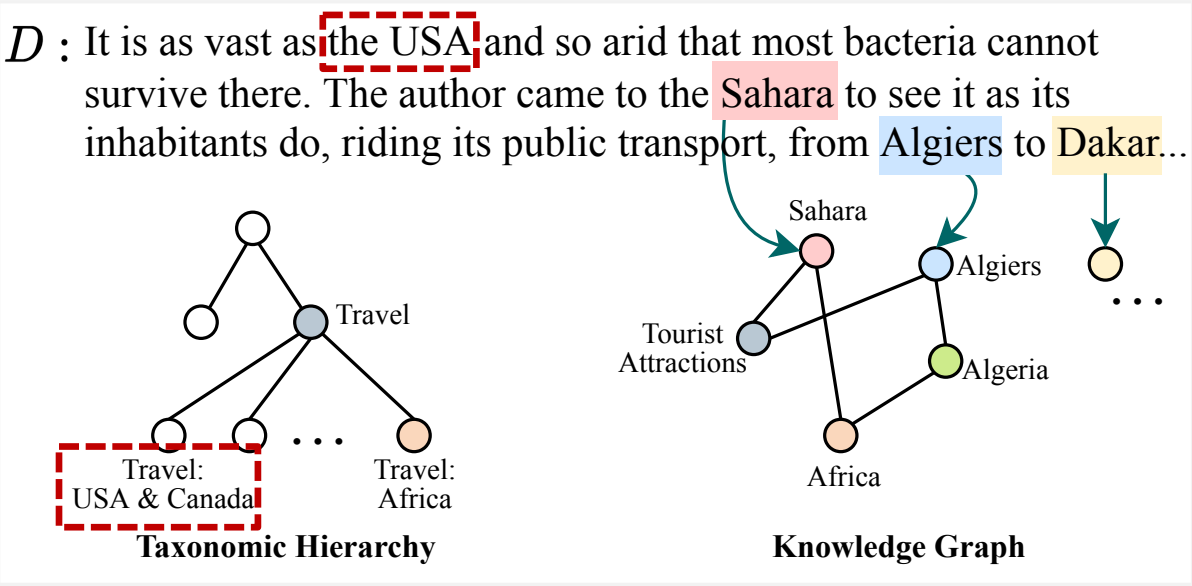
知识概念嵌入增强的层级文本分类方法

- 相关工作
  - **局部层级文本分类方法** [1,2]
    - ✓ 为每个局部区域（如：每个类别或每个层级）训练一个单独的分类器
  - **全局层级文本分类方法** [3,4,5]
    - ✓ 为所有类别构建一个单一的分类器
- 缺陷
  - 仅从输入文本和标签结构中进行表示学习，模型 **缺乏领域知识**
  - 面对涉及深层语义的问题时，模型表现较差

[1] Siddhartha Banerjee. Hierarchical transfer learning for multi-label text classification. ACL-2019.
[2] Kazuya Shimura. Hft-cnn: Learning hierarchical category structure for multi-label short text categorization. EMNLP-2018.
[3] Jie Zhou. Hierarchy-aware global model for hierarchical text classification. ACL-2020.
[4] Haibin Chen. Hierarchy-aware label semantics matching network for hierarchical text classification. ACL-2021.
[5] Zihan Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. ACL2022.

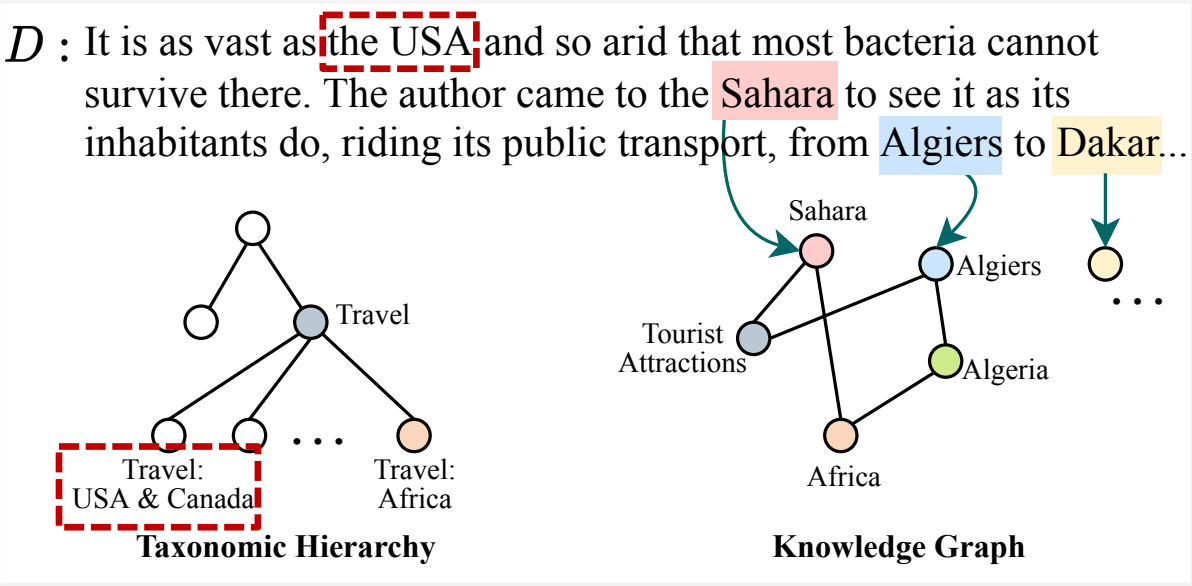
知识概念嵌入增强的层级文本分类方法

- 传统模型
  - The USA → Travel: USA & Canada
- 结合 **知识概念网络**
  - $(Sahara, part\_of, Africa)$, $(Algiers, capital\_of, Algeria)$
  - → Travel、Travel: Africa



知识概念嵌入增强的层级文本分类方法

- 传统模型
  - The USA → Travel: USA & Canada
- 结合 **知识概念网络**
  - $(Sahara, part\_of, Africa)$, $(Algiers, capital\_of, Algeria)$
  - → Travel、Travel: Africa



知识概念嵌入增强的层级文本分类方法

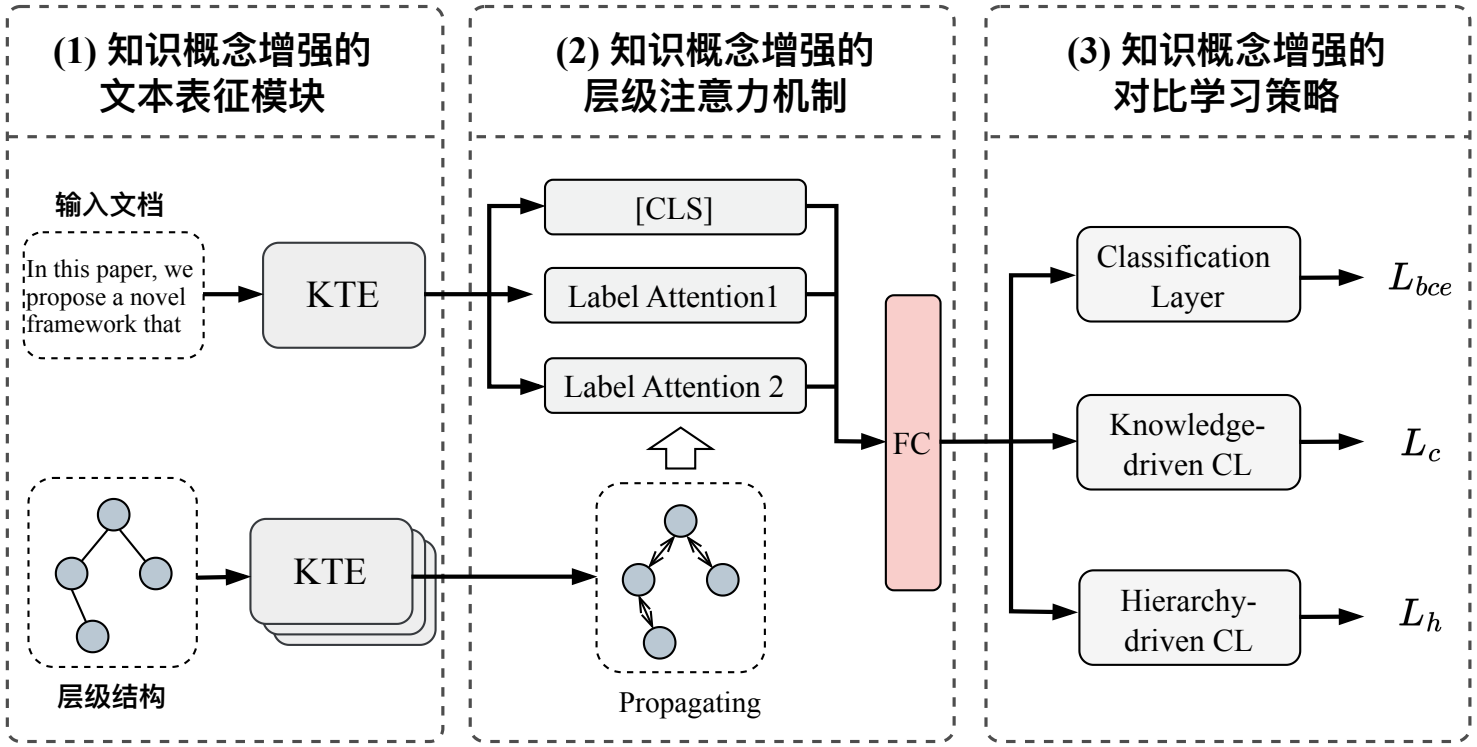
- 传统模型
  - The USA → Travel: USA & Canada

- 结合 **知识概念网络**
  - $(Sahara, part\_of, Africa)$, $(Algiers, capital\_of, Algeria)$
  - → Travel、Travel: Africa

如何将知识概念网络中的知识融入到层级文本分类过程中，从而缓解领域知识缺乏问题?

知识概念嵌入增强的层级文本分类方法

- 将 **知识概念网络** 创新性地融入 **文本表征**、**层级标签学习** 以及 **模型的训练策略** 之中
  - 构建了一个能够充分利用 **外部知识** 的 **层级文本分类方法**



知识概念嵌入增强的层级文本分类方法

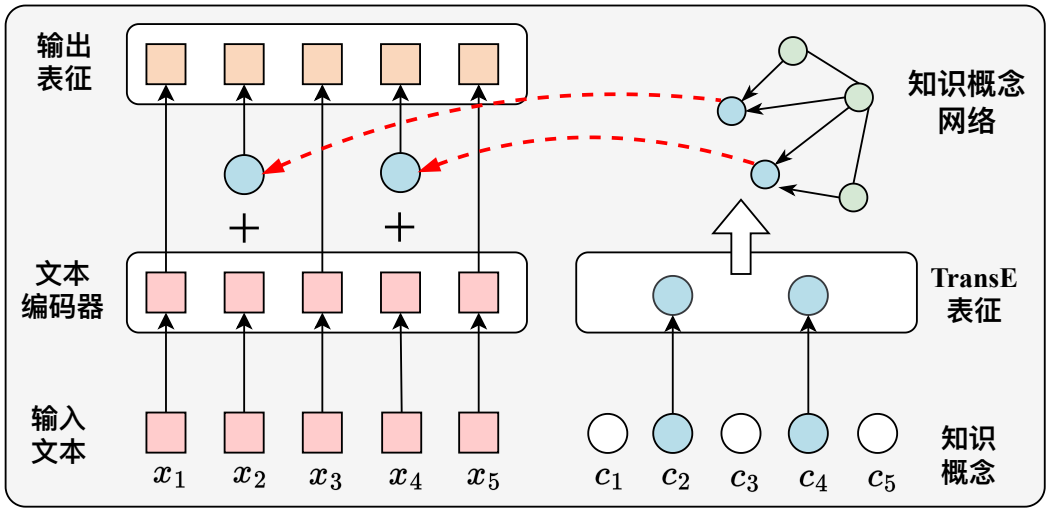
- **知识概念增强文本编码器**
  - 将知识概念网络结合进输入文本的表征学习之中
  - **语义表征** + **结构化知识表征**

$$\{w_1, ..., w_N\} = BERT(\{x_1, ..., x_N\})$$

$$\{u_1, ..., u_N\} = U(\{c_1, ..., c_N\})$$

$$\{m_1, ..., m_N\} = \{w_1 + u'_1, ..., w_N + u'_N\}$$

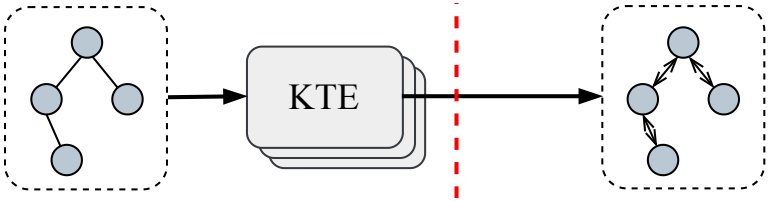
文本语义表征



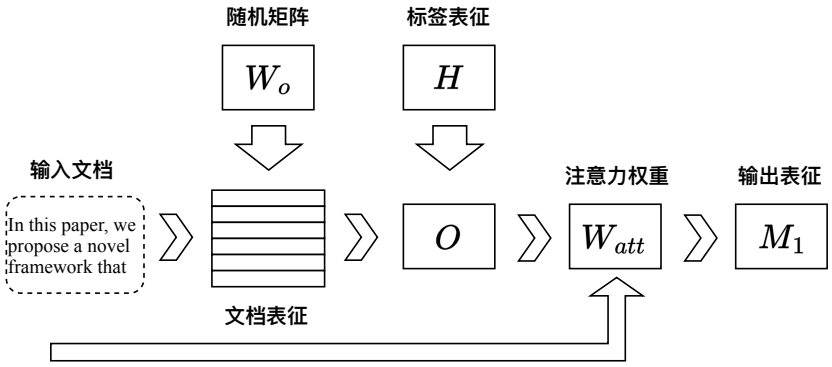
知识概念网络表征

知识概念嵌入增强的层级文本分类方法

- **知识概念增强的层级注意力机制**
- **标签表征学习**
  - 知识感知的标签语义表征 → 基于层级结构进行传播优化
- **标签注意力机制**
  - 通过标签注意力机制 **增强文档表征**



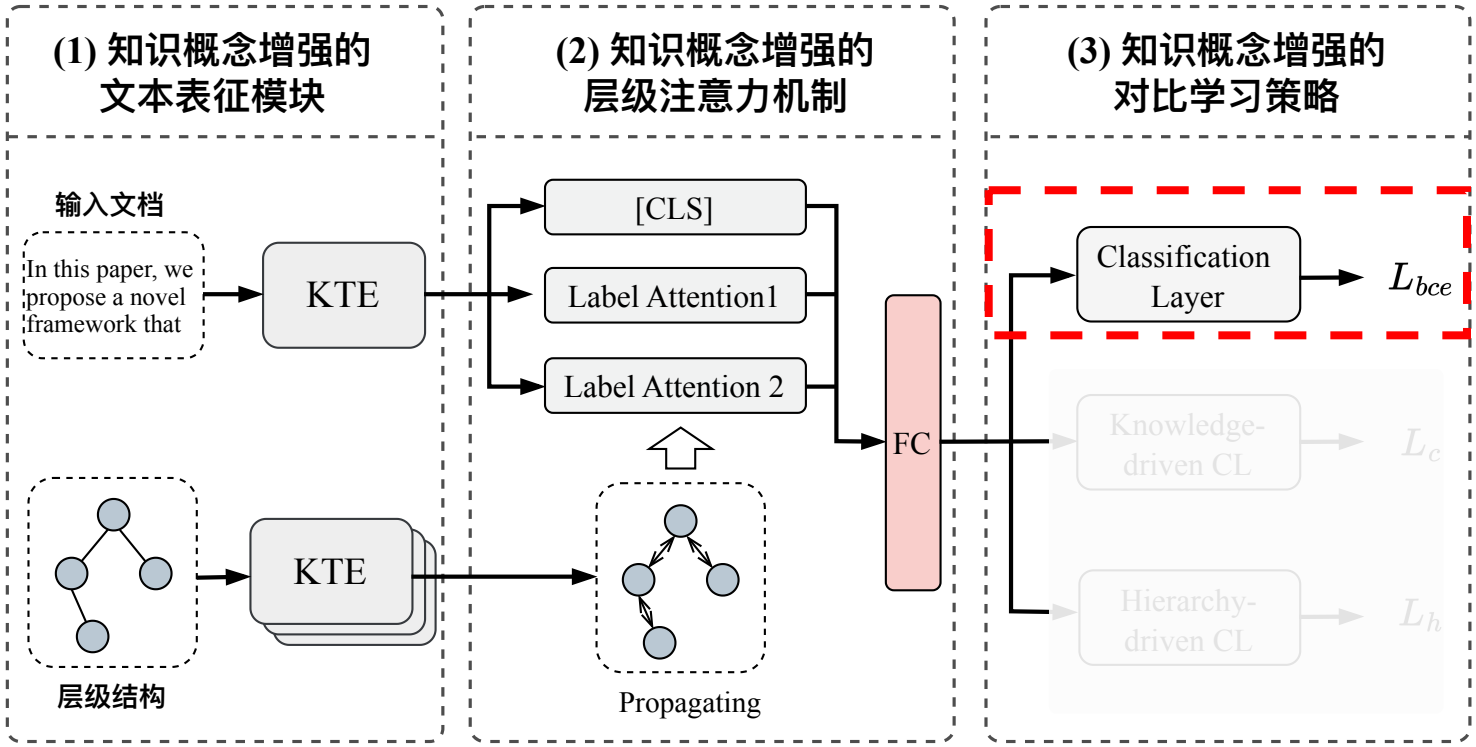
$$R_l^i = mean(KTE(L_i)), i = 1, ..., K,$$
$$R_l = [R_l^1, R_l^2, ..., R_l^K],$$
$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}),$$



$$R_d = KTE(D),$$
$$O = tanh(W_o \cdot R_d^T),$$
$$W_{att} = softmax(H \cdot O),$$
$$M_1 = mean(W_{att} \cdot R_d),$$

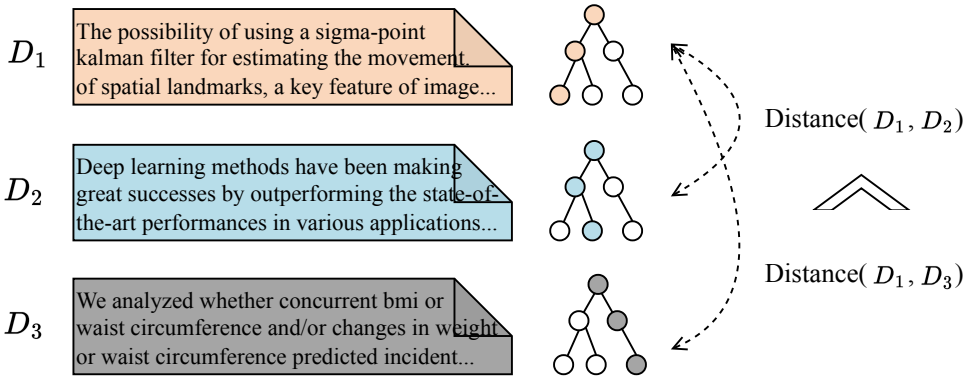
知识概念嵌入增强的层级文本分类方法

- 将 **知识概念网络** 创新性地融入 **文本表征**、**层级标签学习** 以及 **模型的训练策略** 之中
  - 构建了一个能够充分利用 **外部知识** 的 **层级文本分类方法**



知识概念嵌入增强的层级文本分类方法

- **知识概念增强的对比学习策略**
  - 层级标签体系下，**共同标签越多** 的文档的表征越接近
  - 结合知识概念网络，随着层级的加深，**共享知识概念** 的数量逐渐增加，**共享知识概念越多** 的文档的表征也越接近
  - → **渐进式的距离关系** ***Progressive Distance Relationship***



共享标签示例

层级	BGC 数据集	WOS 数据集
L-1	4.29	5.82
L-2	4.93	8.00
L-3	5.96	–
L-4	5.94	–
Total	3.12	4.87

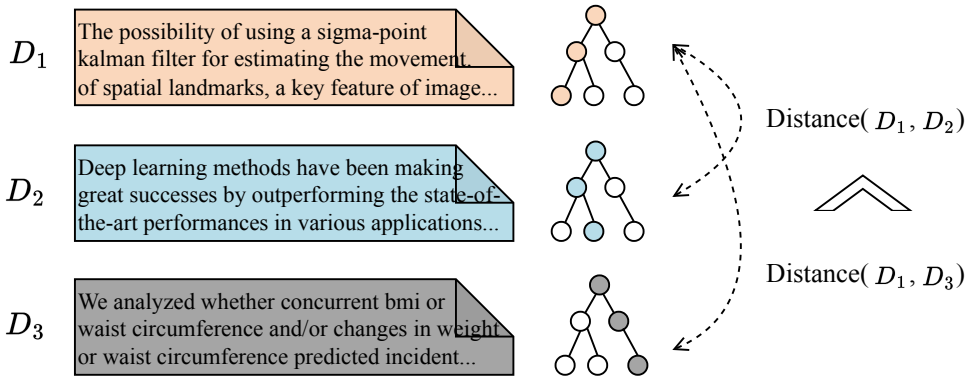
共享知识概念分析

知识概念嵌入增强的层级文本分类方法

- **知识概念增强的对比学习策略**
  - → 渐进式的距离关系 *Progressive Distance Relationship*
  - **知识 / 层级** 驱动的对比学习
  - 在一个大小为 b 的批次中，定义如下 Loss：

$$L_c^{ij} = -\beta_{ij} \log \frac{e^{-d(z_i,z_j)/\tau}}{\sum_{k\in g(i)} e^{-d(z_i,z_k)/\tau}},$$

$$c_{ij} = |C_i \cap C_j|, \quad \beta_{ij} = \frac{c_{ij}}{\sum_{k\in g(i)} c_{ik}},$$

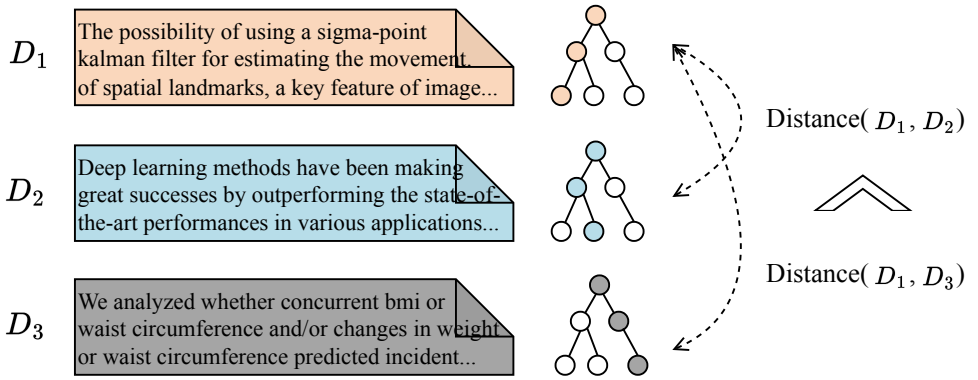


知识概念嵌入增强的层级文本分类方法

- **知识概念增强的对比学习策略**
  - → 渐进式的距离关系 *Progressive Distance Relationship*
  - **知识 / 层级** 驱动的对比学习
  - 在一个大小为 b 的批次中，定义如下 Loss：

$$\downarrow L_c^{ij} = -\beta_{ij} \log \frac{e^{-d(z_i,z_j)/\tau}}{\sum_{k\in g(i)} e^{-d(z_i,z_k)/\tau}},$$

$$\uparrow c_{ij} = |C_i \cap C_j|, \quad \uparrow \beta_{ij} = \frac{c_{ij}}{\sum_{k\in g(i)} c_{ik}},$$



知识概念嵌入增强的层级文本分类方法

- **知识概念增强的对比学习策略**
  - → **渐进式的距离关系** *Progressive Distance Relationship*
  - **知识 / 层级** 驱动的对比学习
  - 在一个大小为 b 的批次中，定义如下 Loss：

$$L = L_{bce} + \lambda_c L_c + \lambda_h L_h$$

与分类损失函数结合，进行高效训练

知识概念嵌入增强的层级文本分类方法

实验数据集

- BlurbGenreCollection-EN (BGC)
  - 主要包含书籍的广告描述
- Web-of-Science (WOS)
  - 包含来自 Web of Science 的已发表的论文

统计指标	BGC	WOS
类别数量	146	141
层级数量	4	2
平均每个样本的标签数量	3.01	2.0
训练集	58,715	30,070
验证集	14,785	7,518
测试集	18,394	9,397

知识概念嵌入增强的层级文本分类方法

实验结果

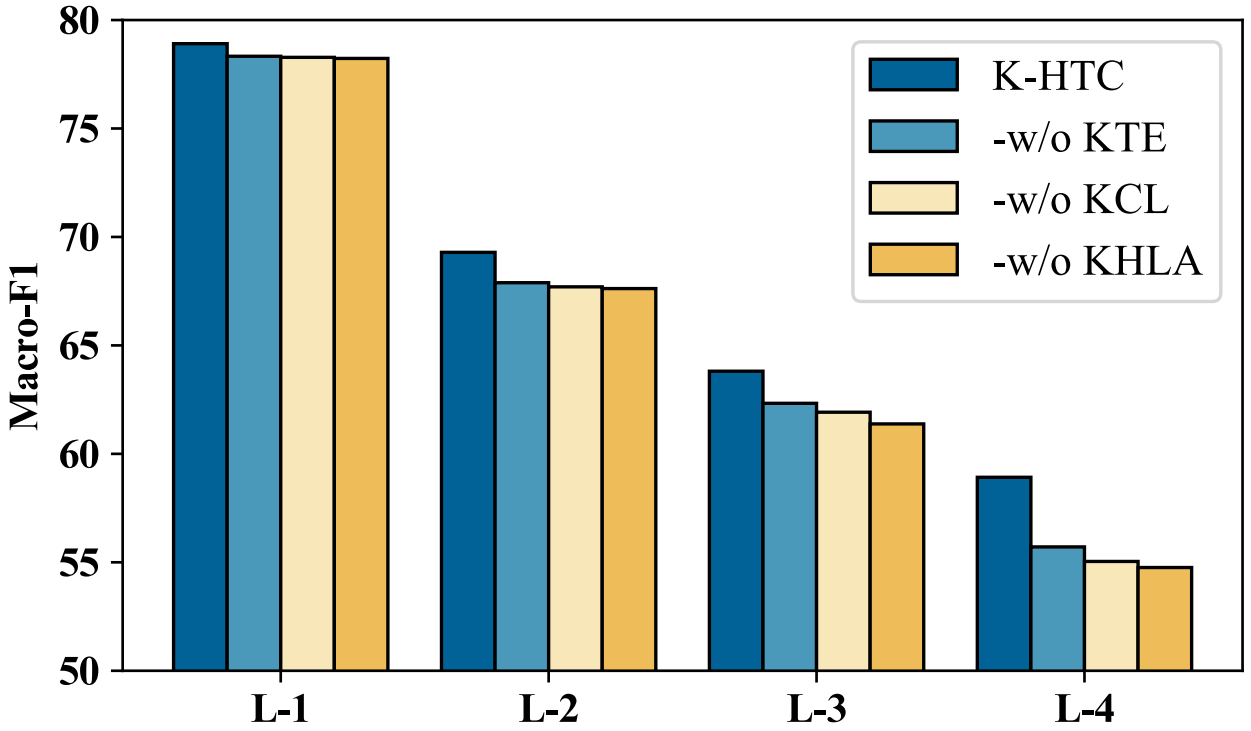
- 本研究提出的 **K-HTC 方法** 在绝大部分指标上均优于所有基线方法

Methods	BGC				WOS			
	Pre	Rec	Ma-F1	Mi-F1	Pre	Rec	Ma-F1	Mi-F1
Hierarchy-Aware Methods								
HiAGM	57.41	53.45	54.71	74.49	82.77	78.12	80.05	85.95
HTCInfoMax	61.58	52.38	55.18	73.52	80.90	77.27	78.64	84.65
HiMatch	59.50	52.88	55.08	74.98	83.26	77.94	80.09	86.04
Pre-trained Language Methods								
HiAGM+BERT	65.61	61.79	62.98	78.62	81.81	78.86	80.09	85.83
HTCInfoMax+BERT	65.47	62.15	62.87	78.47	79.95	79.59	79.33	85.18
HiMatch+BERT	64.67	62.05	62.62	79.23	82.29	80.00	80.92	86.46
KW-BERT	66.39	62.68	63.72	79.24	82.88	78.75	80.30	86.19
HGCLR	67.65	61.28	63.64	79.36	83.67	79.30	81.02	87.01
HPT	70.27	62.70	65.33	**80.72**	83.71	79.74	81.10	86.82
K-HTC (ours)	**71.26**	**63.31**	**65.99**	80.52	**84.15**	**80.01**	**81.69**	**87.29**

知识概念嵌入增强的层级文本分类方法

不同层级上的表现分析

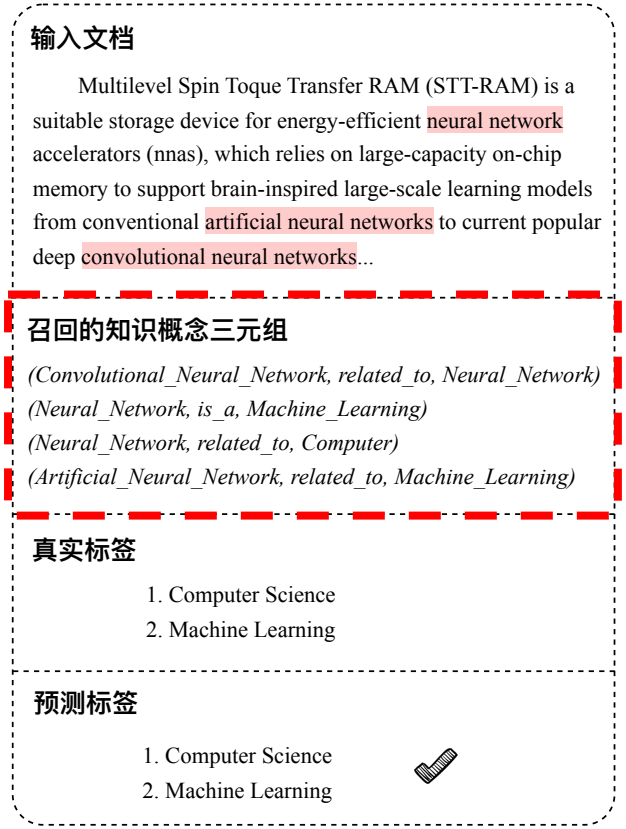
- **随着层级加深**，所有方法的性能均出现了明显的下降
- K-HTC 与其消融变体之间的差距随着 **层级深度的增加而扩大**



知识概念嵌入增强的层级文本分类方法

案例分析

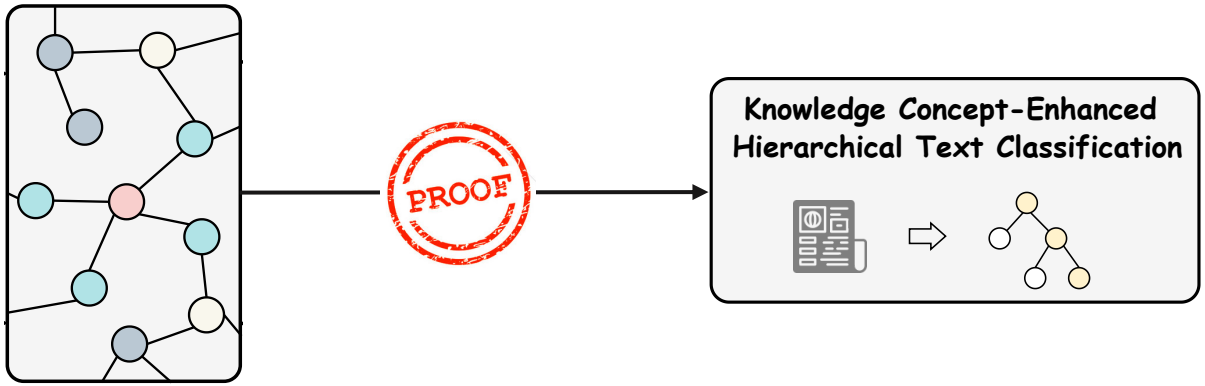
- 借助知识概念网络中的知识，K-HTC 能够合理地得出正确的推断
- 显式的 **知识概念三元组**，这也为预测结果提供了更多的 **可解释性空间**，进一步促进了 **可信的层级文本分类**



知识概念嵌入增强的层级文本分类方法

本章小结

- 将所构建的 **知识概念网络** 应用在 **层级文本分类** 任务上
- 提出了 **知识概念嵌入增强的层级文本分类方法**
- 证明了 **知识概念挖掘与应用** 对 **可信的层级文本分类** 有显著的提升作用



相关工作以第一作者身份发表于 CCF A 类会议

ACL2023 上

知识概念嵌入增强的层级文本分类方法

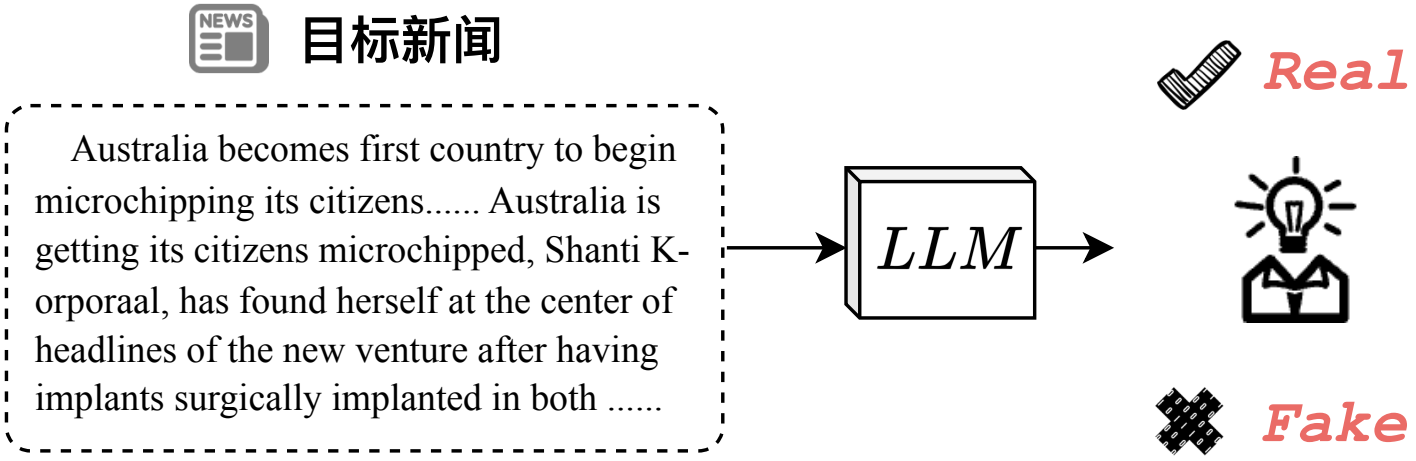
- 围绕 **知识概念的抽取**、**关联** 与 **应用** 开展技术研究
  - 形成 **基于层级约束和语义建模的无监督知识概念抽取方法** -- **交互式增强的小样本知识概念关联方法** -- **知识概念嵌入增强的层级文本分类方法** -- **知识概念引导的虚假新闻检测方法** 的研究框架

汇报提纲

1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. 知识概念嵌入增强的层级文本分类方法
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

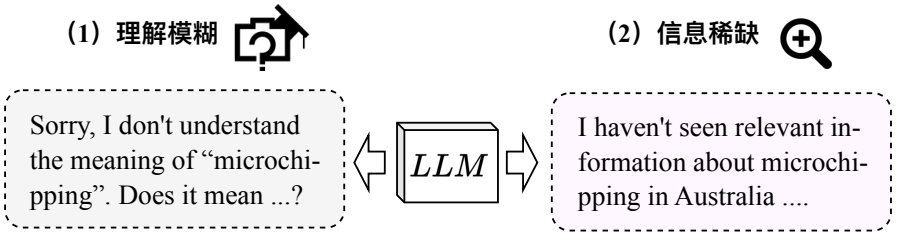
知识概念引导的虚假新闻检测方法

- **虚假新闻检测**，旨在区分虚假新闻与真实新闻
- 伴随着互联网的快速发展，虚假、谣言新闻等在社交媒体平台上广泛传播，如何在**低数据资源场景下**，高效、准确的识别虚假新闻正在受到越来越多研究者的关注和重视



知识概念引导的虚假新闻检测方法

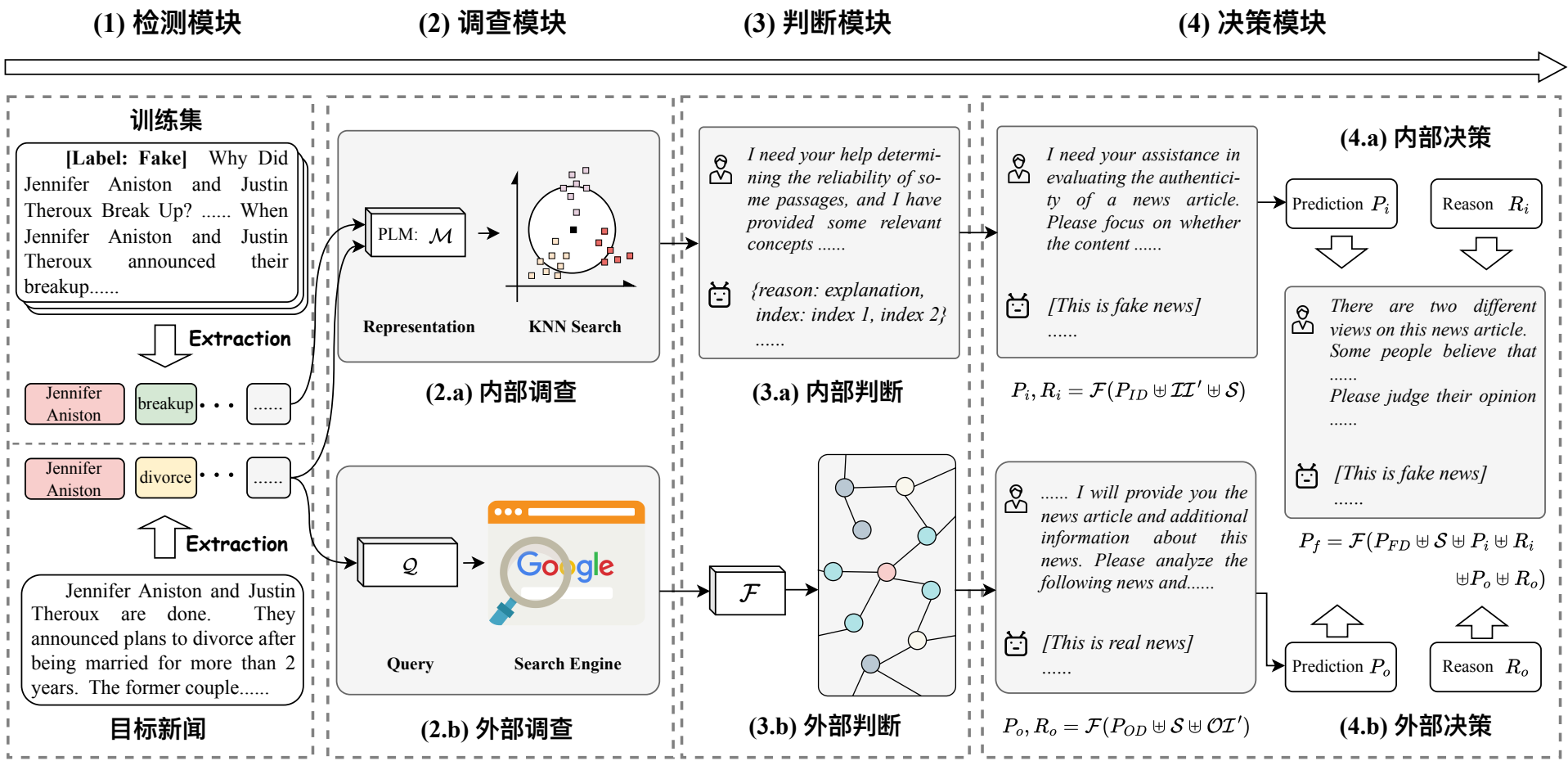
- 现有虚假新闻检测方法
  - 传统方法：FakeFlow（EACL’ 2021）
    - ✓ 需要 **大量标注数据** 进行分类训练，**小样本场景** 下表现差
  - 大语言模型的 **上下文学习方法**：ARG（ AAAI’ 2024）
    - ✓ 依赖人工设计的提示指令，直接询问大语言模型给定新闻的真实性



- 大语言模型方法的缺陷
  - **理解模糊性**
    - ✓ 遇到小领域内的相关新闻时，难以理解其核心内容的含义
  - **信息稀缺性**
    - ✓ 新闻语料具有 **动态性** 和 **快速更新** 的特点，而大语言模型的训练语料往往是 **过时** 的，这导致其在检测过程中缺乏最新的相关信息。

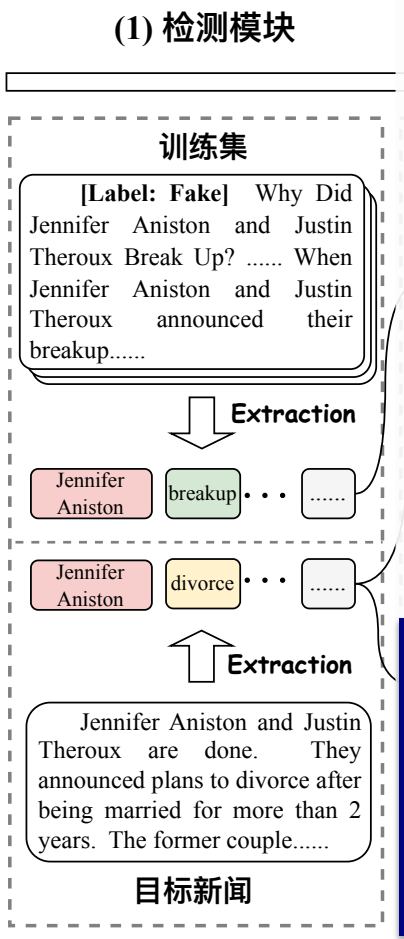
知识概念引导的虚假新闻检测方法

- 在 **知识概念网络的引导下**，从 **内、外两个视角** 进行建模：
- （1）检测模块；（2）调查模块；（3）判断模块；（4）决策模块



知识概念引导的虚假新闻检测方法

- **检测模块**
  - 识别新闻文本中的 **知识概念**，提取重要的 **关键知识概念集合**



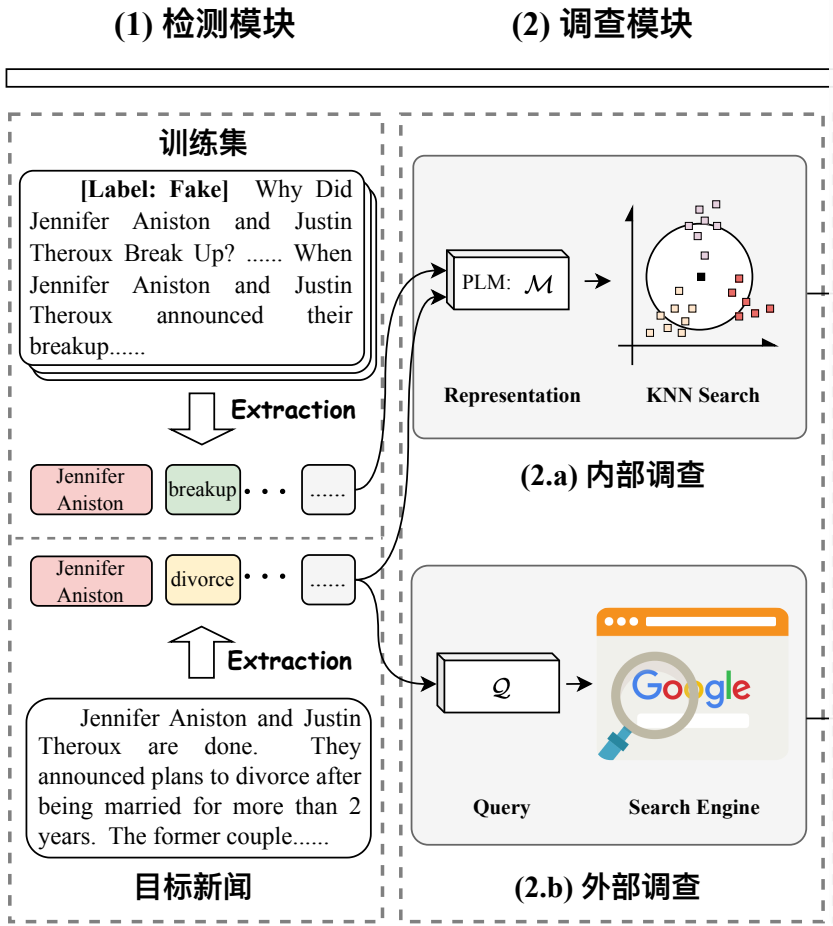
$$\{c_1, c_2, \ldots, c_m\} = \mathcal{E}(S)$$

$$C' = \mathcal{F}(P_{detection}, S, C)$$

How relevant these knowledge concepts are to the given news?

知识概念引导的虚假新闻检测方法

- **调查模块**
  - **内部**：训练集合；**外部**：Google搜索引擎



内部视角：训练集合

$$\mathcal{II} = \{U_{positive}, U_{negative}\}$$

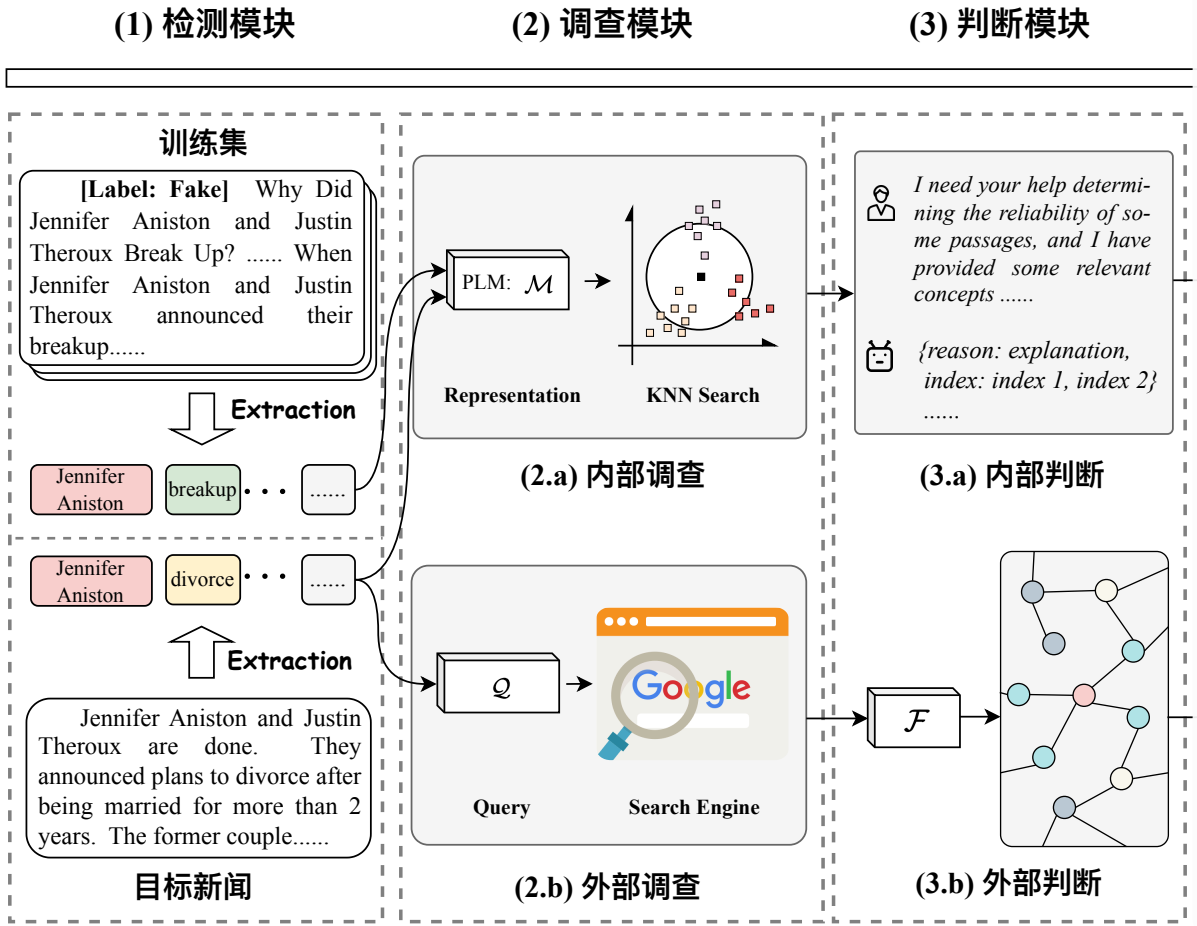
外部视角：Google搜索引擎

$$\mathcal{OI} = \{G_k,\ k = 1, 2, ...L\}$$

知识概念引导的虚假新闻检测方法

判断模块

- **内部**：判断相关性；**外部**：借助知识概念网络判断真实性



内部视角： 判断相关性

$$U'_{positive} = \mathcal{F}(P_{IJ} \uplus U_{positive} \uplus C')$$

$$U'_{negative} = \mathcal{F}(P_{IJ} \uplus U_{negative} \uplus C')$$

外部视角： 知识网络判断真实性

$$f(h,r,t) = \mathrm{Re}\,(e_h \cdot r \cdot \overline{e_t})$$

$$Sc(T_k^l) = f(h_k^l, r_k^l, t_k^l)$$

知识概念引导的虚假新闻检测方法

- **决策模块**
  - **内部决策** -- **外部决策** -- **综合决策**

内部视角： 基于内部信息预测

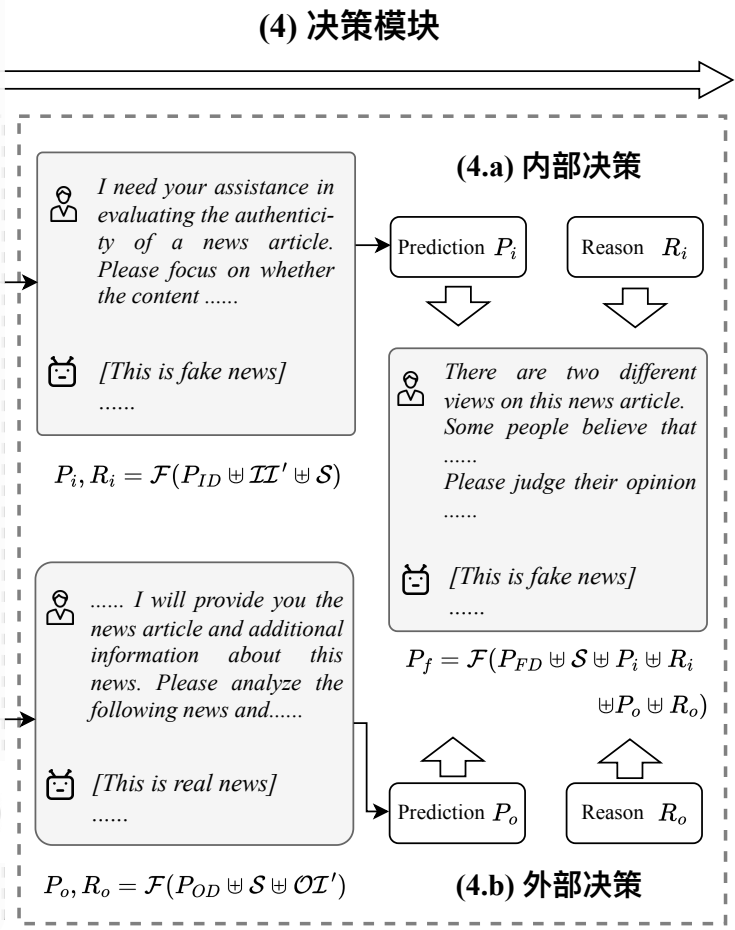
$$P_i,\ R_i = \mathcal{F}(P_{ID} \uplus \mathcal{II}' \uplus S)$$

外部视角： 基于外部检索信息预测

$$P_o,\ R_o = \mathcal{F}(P_{OD} \uplus S \uplus \mathcal{OI}')$$

综合视角： 综合考虑内、外部两个视角

$$P_f = \mathcal{F}(P_{FD} \uplus S \uplus P_i \uplus R_i \uplus P_o \uplus R_o)$$



知识概念引导的虚假新闻检测方法

实验数据集

- **FakeNewsNet** 基准库
  - **PolitiFact**：政治新闻检测数据集
  - **Gossipcop**：娱乐新闻事实核查数据集
  - 随机选择 $K \in (8, 32, 100)$ 的正负新闻样本作为训练集

Dataset		PolitiFact	Gossipcop
Train	# True news	8/32/100	8/32/100
	# Fake news	8/32/100	8/32/100
	# Total news	16/64/200	16/64/200
Test	# True news	120	3,200
	# Fake news	80	1,060
	# Total news	200	4,260

知识概念引导的虚假新闻检测方法

实验结果

- 本研究提出的 DAFND 模型在各项指标上均超越了所有对比方法
- 尤其在 **训练数据受限**（K 更小）时，**领先优势** 更明显

Dataset	Methods	ACC			F-1 score		
		K=8	K=32	K=100	K=8	K=32	K=100
PolitiFact	① PROPANEWS	40.00	43.50	40.00	57.14	58.30	57.14
	② FakeFlow	61.00	62.50	63.50	44.29	47.55	48.95
	③ MDFEND	65.50	64.00	71.50	62.30	64.36	69.84
	④ PSM	70.00	72.50	79.00	49.15	52.38	65.70
	⑤ KPL	58.33	73.44	82.29	60.40	73.58	81.11
	⑥ Auto-CoT	49.50	58.00	64.00	53.88	58.00	55.00
	⑦ Zephyr	60.00	63.50	66.50	48.72	53.50	54.42
	⑧ ChatGLM-3	68.50	68.50	72.50	58.28	58.82	64.05
	⑨ LLama-3	69.50	70.50	69.00	63.91	65.09	64.00
	⑩ GPT-3.5	71.00	69.50	73.00	60.27	60.65	64.47
	⑪ ARG	74.00	78.50	82.50	67.16	68.61	80.61
	DAFND (ours)	**87.00**	**88.00**	**89.00**	**82.43**	**83.78**	**85.33**

知识概念引导的虚假新闻检测方法

消融实验分析

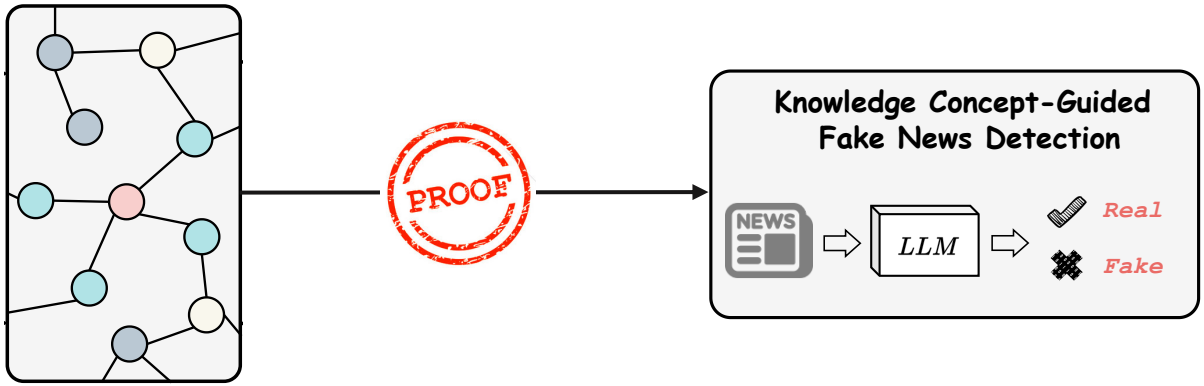
- 从 **内、外两个视角** 出发，得到以下 3 种消融变体：
  - DAFND - Inside：简化内部视角
  - DAFND - Outside：简化外部视角
  - DAFND – Both：从内外部两个视角简化
  - Zephyr：最终简化版本

Methods	ACC	F-1 score
DAFND	**89.00**	**85.33**
DAFND - Inside	86.50	81.38
DAFND - Outside	86.00	81.58
DAFND - Both	85.00	79.45
⑦ Zephyr	66.50	54.42

知识概念引导的虚假新闻检测方法

本章小结

- 将 **知识概念网络** 应用在虚假新闻检测任务上
- 提出了 **知识概念引导的虚假新闻检测方法**
- 证明了 **知识概念挖掘与应用** 对 **可信的虚假新闻检测** 有显著提升作用



相关工作以第一作者身份投稿于 CCF A 类会议

ACL2025 上

汇报提纲

1. 研究背景
2. 基于层级约束和语义建模的无监督知识概念抽取方法
3. 交互式增强的小样本知识概念关联方法
4. 知识概念嵌入增强的层级文本分类方法
5. 知识概念引导的虚假新闻检测方法
6. 总结与展望

总结：研究成果

面向可信自然语言处理的知识概念挖掘与应用方法研究

研究背景：现有的自然语言处理技术存在**领域知识缺乏**、**可信度低**、**可解释性差**等问题，难以构建**可信的自然语言处理算法**

“面向可信自然语言处理的知识概念挖掘与应用方法研究”的**若干挑战**

- 知识概念语义复杂 **解析识别难**
- 知识概念分布分散 **关联建模难**
- 知识概念场景多样 **下游应用难**

- **研究内容一**：*基于层级约束和语义建模的无监督知识概念抽取方法*
- 数据基础 →
- **研究内容二**：*交互式增强的小样本知识概念关联方法*
- 相关知识 →
  - **研究内容三**：*知识概念嵌入增强的层级文本分类方法*
  - **研究内容四**：*知识概念引导的虚假新闻检测方法*

可信的知识概念网络构建 → *可信的自然语言处理应用*

总体目标

分析

研究挑战

支持

算法框架

构建 → 应用

总结：研究成果

学术成果

发表学术论文超过 15 篇（CCF A 类 6 篇，B 类 9 篇）

- 1 篇论文获 CICAI 会议 **“Finalist of Best Paper Award”** 奖项（最佳论文季军）

第一作者 论文 4 篇（CCF A 类 1 篇，B 类 3 篇）

- **ACL2023**、**ICDM2020**、**TKDD2023**、**DASFAA2024**
- 另有 1 篇第一作者论文投稿至 **ACL2025**， 1 篇共同一作论文投稿至 **KDD2025**

参与项目

■ 2022 年 – 2023 年	基于百科文本的知识图谱构建	校企合作项目 - 字节跳动（结题）
■ 2021 年 – 2025 年	面向终身学习的个性化“数字教师”智能体技术研究与应用	国家重点研发计划项目（在研）
■ 2021 年 – 2025 年	基于多模态数据的学习者认知诊断理论与关键技术研究	国家自然科学基金联合基金项目（在研）

一作论文列表

一作论文列表

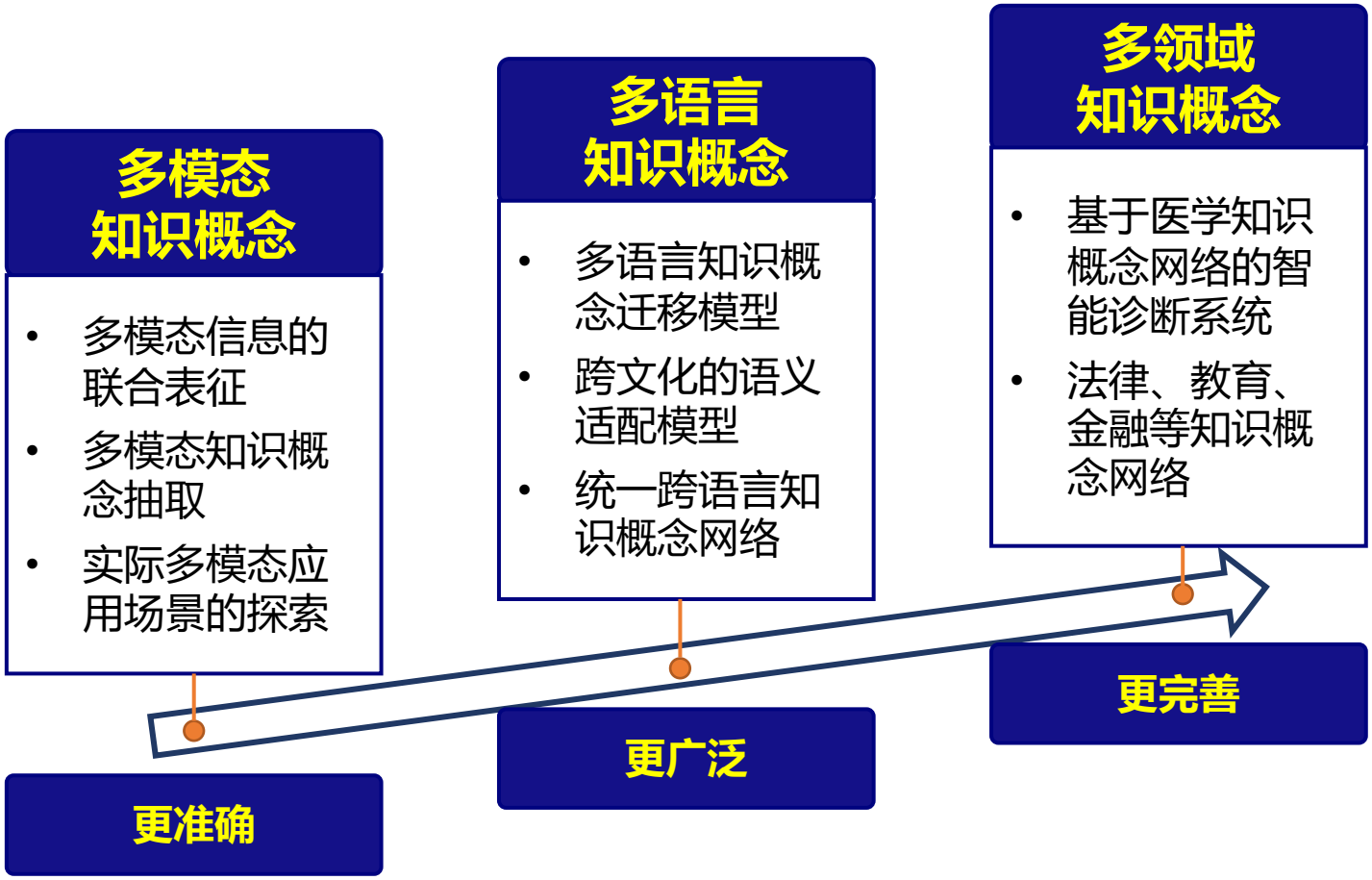
1. **Ye Liu**, Kai Zhang*, Zhenya Huang, Kehang Wang, Yanghai Zhang, Qi Liu, Enhong Chen*. Enhancing Hierarchical Text Classification through Knowledge Graph Integration. Findings of the 61st annual meeting of the Association for Computational Linguistics (ACL), 2023. **(CCF A)**
2. **Ye Liu**, Han Wu, Zhenya Huang, Hao Wang, Jianhui Ma, Qi Liu, Enhong Chen*, et al. Technical Phrase Extraction for Patent Mining: A Multi-level Approach. The 2020 IEEE International Conference on Data Mining (ICDM), 2020. **(CCF B)**
3. **Ye Liu**, Han Wu, Zhenya Huang, Hao Wang, Yuting Ning, Jianhui Ma, Qi Liu, Enhong Chen*. TechPat: Technical Phrase Extraction for Patent Mining. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 2023. **(CCF B)**
4. **Ye Liu**, Kai Zhang*, Aoran Gan, Linan Yue, Feng Hu, Qi Liu, Enhong Chen. Empowering Few-Shot Relation Extraction with The Integration of Traditional RE Methods and Large Language Models. The 29th International Conference on Database Systems for Advanced Applications (DASFAA), 2024. **(CCF B)**
5. **Ye Liu**, et al. Detect, Investigate, Judge and Determine: A Knowledge-guided Framework for Few-shot Fake News Detection. Submitted to ACL2025. **(CCF A)**
6. Haoyu Tang†, **Ye Liu†**, et al. Learn while Unlearn: An Iterative Unlearning Framework for Generative Language Models. Submitted to KDD2025. **(CCF A)**

专家意见修改

专家意见修改

- 论文各章的框架图应采用统一的模式
- 关键术语给出清晰的定义或准确说明
- 给出评价指标的定义和具体计算公式
- 细化可信自然语言处理的解释
- 报告消融实验的具体数值
- 统一参考文献的格式
- 重新修改了摘要
- 文字润色
- ……

未来展望



从**多模态**、**多语言**、**多领域**的需求出发，进一步提升模型的**泛化能力**、**动态适应性**和**实际应用价值**，为构建更加**智能化**、**可信赖**的自然语言处理、人工智能系统提供坚实的技术支持。

致谢

陈恩红 教授

周晓方 教授

感谢各位专家！

敬请指正！