CCL25-Eval任务四系统报告：基于RAG与谓词相似性方法的叙实性检测智能体

王昱[1]，杨倩[2]，梁科[1]，杨怿恒[1]，翟雨[1]，黄居仁[1]
[1]香港理工大学/ 香港
[2]四川大学/ 四川
janet-yu.wang@connect.polyu.hk
yangqian.official@gmail.com
{leo-ke.liang, yiheng.yang, tonyayu.zhai}@connect.polyu.hk
churen.huang@polyu.edu.hk

摘要

本文聚焦于“叙实性推理”任务，即判断语言中事件真实性的语义理解能力。该任务不依赖外部知识，而基于语言结构本身进行推理，对当前大语言模型（LLMs）提出挑战。为解决模型在叙实性漂移、多义词处理等方面的不足，作者提出一种结合RAG（检索增强生成）与谓词相似性的方法，构建了一个融合参数化与非参数化知识的叙实性检测智能体系统。该系统通过分步提示与知识库支持，实现了更高的一致性、准确性与可解释性，在评测任务中取得了0.9240的稳健表现。

关键词： 叙实性 ； RAG ； 谓词相似性 ； Agent

System Report for CCL25-Eval Task 4: A Factivity Detection Agent Based on RAG and Predicate Similarity Methods

Yu Wang[1], Qian Yang[2], Ke Liang[1], Yiheng Yang[1], Yu Zhai[1], Churen Huang[1]
[1]The Hong Kong Polytechnic University / Hong Kong
[2]Sichuan University / Sichuan
janet-yu.wang@connect.polyu.hk
yangqian.official@gmail.com
{leo-ke.liang, yiheng.yang, tonyayu.zhai}@connect.polyu.hk
churen.huang@polyu.edu.hk

Abstract

This paper addresses factivity inference—determining event truth based on linguistic cues. To improve large language models' performance on this task, the authors propose an agent system combining Retrieval-Augmented Generation (RAG) and predicate similarity. The method integrates structured and contextual knowledge, enabling accurate, consistent, and interpretable truth-value judgments. It achieves a strong evaluation score of 0.9240.

Keywords: Factivity , RAG , Predicate similarity , Agent

1 引言

叙实性推理(Factivity Inference, FI)是一种聚焦事件真实性判断的语义理解任务(Kiparsky and Kiparsky, 1970)。在日常对话中，拥有叙实性推理能力的语言使用者往往可以通过某些动词性语言成分（例如“知道”、“谎称”、“猜测”等）推测说话人对所述事件真伪情况的态度。例如，无论是“他知道会议已经开始了”还是“他不知道会议已经开始了”，尽管句子本身并未直接

陈述“会议开始”这一事件的真实性，具备叙实性推理能力的听话者仍会自然地理解该事件已然发生，属于事实。这种基于叙实性的推理并非依赖于世界知识(World Knowledge)，而是建立在语言使用者内化的分析性语言知识(Analytical Knowledge of Language) 之上。例如，在上述句子中，无论是“知道”还是“不知道”，听话者无需查证具体的时间、地点或会议背景，仅凭“知道”这一动词所携带的语义预设，即可推断“会议开始”这一事件为真。因此，这种不依赖外部知识、而是依托语言结构本身进行的推理机制，对当前主要依赖知识检索与共现数据(Co-occurrence Data)进行事实判断的大语言模型(Large Language Models, LLMs) 提出了挑战。

学者们将汉语中的叙实性动词分为三大类：正叙实动词、反叙实动词和非叙实动词(袁毓林, 2014; 李新良and 袁毓林, 2016; 李新良, 2020; 李新良et al., 2023; 王昱, 2024)。具体来说，肯定式和否定式都预设其宾语小句为真的动词是正叙实动词（如“知道”、“意识到”），肯定式和否定式都不预设其宾语小句为真，也不预设其宾语小句为假的动词是非叙实动词（如“猜测”、“认为”），肯定式和否定式都预设其宾语小句为假的动词是反叙实动词（如“谎称”、“污蔑”）。而随着对中文真实语料的深入分析，学者们注意到，叙实性动词的真值预设并非恒定，而是会根据语境发生变化，这一现象被称为叙实性漂移(李新良, 2018; 王昱and 袁毓林, 2020; 袁毓林, 2020a; 袁毓林, 2020b; Wang and Yuan, 2021)。即便是同一个谓词，在不同语境下，其所暗示的事实立场也可能截然不同。例如，“她记得那次旅行很愉快”通常意味着“那次旅行很愉快”这一命题为真；但“她记得自己可能没去成那家餐馆”中，则包含较强的不确定性，语义上更接近推测而非事实陈述(袁毓林, 2020b)。又如，“有人哀叹物价飞涨”可能更多是一种情绪表达，未必表示说话者相信物价确实上涨；但若说“专家哀叹市场失灵”，则更可能传达出其对“市场失灵”为事实的认同。这类语义差异，不仅取决于谓词本身，也与说话者权威性和上下文语境密切相关(袁毓林, 2020c)。因此，仅靠列出的“正叙实”“反叙实”或“非叙实”标签，并不足以应对实际语言中多义、多变的语用现象。我们亟需一种能够灵活适应语境变化、动态整合知识与语义推理的机制。

叙实性推理是大模型文本蕴涵识别(Textual Entailment Recognizing)、幻觉处理(Hallucination Solving)、信念修正(Belief Revision)等任务的重要语义基础和形式依据。然而，当前的大语言模型在处理此类语义推理任务时表现仍不理想。例如，在评测任务方基于Qwen2-7B-Instruct 所进行的基线测试中，模型在人造语料上的准确率仅为53.74%，常误将推测、虚假陈述或主观判断当作客观事实。这种能力缺失不仅削弱了模型的语篇理解，也影响信息抽取、事实核查和对话系统的效果。能否识别话语中的事实性信息及说话人对事件真实性的态度，对智能体的推理能力和人机交互至关重要。因此，叙实性推理既是衡量语义理解深度的重要指标，也是当前限制语言模型自然推理能力的关键瓶颈之一[1]。

为此，本文提出一种基于RAG (Retrieval-Augmented Generation) (Lewis et al., 2020) 与谓词相似性方法的叙实性检测模型。该方法将参数化知识(Parameterized Knowledge) 与非参数化知识(Non-parameterized Knowledge)相结合，在分类体系基础上引入可动态更新的知识库检索模块，以适应新谓词和语境变体。我们设计了分步变量化提示，引导模型依次完成谓词分类检索、漂移条件获取、语境分析谓词替换与真值推理，从而在提升准确性与覆盖率的同时，增强模型的可解释性。

2 系统方法

通过大模型对语料进行批量真实性检测主要有以下两个难点：

- **从大模型表现来说**：如何保持输出的一致性，前后标准一致；
- **从叙实性任务来说**：
  (1) **叙实性漂移**；在自然语料中，叙实性会随着语境的变化而发生转移。例如“哀叹”是正叙实动词，但其叙实性可被语境取消。
    a. “有人哀叹中国电影没救了”——这是某些人的观点，主体权威性低，并不代表后面的命题为真，因此语境取消了“哀叹”的叙实性，“中国电影没救了”的真值为“U”；
    b. “有人哀叹中国电影没救了，这是不符合事实的”——语境中含有明确否定真值的表达，因此虽然“哀叹”是正叙实动词，它的正叙实性不仅被语境取消，甚至其后命题

[1]https://github.com/UM-FAH-Yuan/FIE2025

的真值在语境中为“F”；

c. “幻想未来自己会变得更美”——“幻想”虽为反叙实动词，但因涉及未来，未来的事情，现在无法判断真假，也许未来“自己真的会变得更美”也可能“自己不会变得更美”，因此反叙实性取消，真值变为“U”。

(2) **多义词**；同一词语在不同语义下具有不同的叙实性，例如“感觉”。

a. “我感觉他不会来”——表达推测，非叙实，他也许会来也许不会来，真值为“U”；

b. “他感觉好痛”——表达自身体验，属于事实，真值为“T”。

2.1 系统方案

本系统采取了RAG的方法构建了叙实性检测的Agent，将参数化知识与非参数化知识相结合。具体分为四个模块：谓词分类检索、漂移条件获取、语境分析谓词替换与真值推理（部分模块名称使用英文，因其为具体任务中的变量名）。

- **谓词分类检索**；我们将谓词按照叙实性分为三个集合，例如反叙实动词是集合b: {想象, 妄称, 污蔑, 诬陷, 装作...}。程序首先判断`predicate`是否在集合中，并给出相应集合中谓词的暂时真值；
- **漂移条件获取**；针对具有特殊叙实性特征、漂移条件的谓词，我们构建了专门的知识库。回答问题时命令模型先在知识库中检索相关信息，再生成回答。此处的知识库是一种外部的非参数化知识，可以随时根据专家意见进行修改，准确性高；
- **语境分析谓词替换**；分析text，结合漂移条件判断truth_value，将其作为暂时真值。对于没有明确漂移条件，叙实性随语境发生变化的谓词，我们采用谓词相似性的方法进行判断，具体方法参照3.1.2。此过程依赖模型内化的参数化知识，即便缺乏相关专家知识也能处理问题，泛化性强；
- **真值推理**；删除text中predicate前的内容，剩余部分为content。删除content和hypothesis中的“确实”，然后分析content中predicate所统辖的内容与hypothesis是否一致或相反：a. 若一致（都是否定或都为肯定），则answer = 暂时真值；b. 若相反（一个是否定，一个是肯定），则answer = 相反的暂时真值。输出最终answer。

3 系统实验

本次评测任务要求参赛队需自选若干大型语言模型，基于测试集数据构造提示词，通过API 向模型逐条发送请求，判断hypothesis在text背景下的真值。数据集规模如下：样例集1000条（人造语料300，真实语料700）；测试集约2000条（人造语料500，真实语料1500）。我们使用扣子平台[2]进行RAG智能体构建，将外部知识库以csv形式注入平台中，模型需要按照步骤，先检索，后生成。实验模型为DeepSeek-V3 (DeepSeek-AI et al., 2024)。通过实验，我们在评测任务上得到0.9240的稳健表现，在人工语料与自然语料上的表现分别为0.9615与0.9099。

3.1 实验结果讨论

3.1.1 难点1：如何确保大模型输出一致

为提升大模型在叙实性判断任务中的一致性与稳定性，我们采用了以下三种策略：

- **变量表达**：在prompt 中通过变量统一指代内容，例如将“删去text中predicate前的内容，留下的文本”定义为content，后续统一使用content表达该部分内容。此方式有助于避免歧义，提升模型理解与输出的一致性。
- **集合表达**：尽管我们引入了基于RAG 的知识库检索机制，但对于大多数叙实性稳定的谓词，仍可通过集合划分（如集合a、b、c）进行归类。这种方式便于模型快速查找并减少判断误差。
- **去除干扰因素**：实验发现，在比对谓词统辖命题与hypothesis命题的一致性时，否定形式及副词（如“确实”）可能引入语义偏差，影响判断结果。因此我们在处理流程中对这类干扰项进行了清理。

通过上述方法优化后，我们在实验中验证了模型输出的一致性达到了100%。

[2]https://www.coze.cn

3.1.2 难点2：如何确保叙实性判断正确

为进一步确保叙实性判断的准确性与鲁棒性，我们采用了以下两项关键方法：

- **谓词替换**：大模型的本质是基于概率的语言建模，即根据已有文本预测最可能的输出，因此，我们可以利用这一特性，基于模型内化的参数化知识，通过谓词相似性替换的方法，对受语境影响的叙实性漂移与多义词进行真值检测。例如“哀叹”，其在知识库中的谓词替换信息如下：
  - 判断方式：请从“认为”、“感到”、“以为”中选择最合适的词替换语境中的“哀叹”：
    * 若替换为“认为”，则truth_value = U；
    * 若替换为“以为”，则truth_value = F；
    * 否则为T。

  通过这个方式，例如“1. 有人哀叹中国电影没救了；2. 有人哀叹中国经济不好，这是不符合事实的；3. 他常常哀叹自己很穷”会被模型替换为“1.有人认为中国电影没救了；2. 有人以为中国经济不好，这是不符合事实的；3. 他常常感到自己很穷。”大模型即可在各种语境影响下合理判断真实性。值得注意的是，选用“认为”“以为”“感到”作为替换词，并非意味着这些词本身不发生叙实性漂移，而是将其作为相似语义的代表。谓词替换实质上是一个语义归类过程，例如替换为“认为”时，模型更倾向于将该语境理解为表达推测。

- **知识库构建**：针对具有特殊叙实性特征的谓词，我们构建了专门的知识库，并通过RAG方式将外部非参数化专家知识注入大模型，增强大模型在叙实性判断上的能力。

	predicate 索引 String	truth_value String	Possible_value String	Possible_Value_Condition String
1	看见	T	U	1. 如果前面有否定形式，构成“没有看见”，t…
2	以为	F	U	如果文中的以为，表述的是误以为，错误认为…
3	哀叹	T	U/F	哀叹的truth_value并不确定，一般为T。 1. 如…
4	怀疑	U	T	一般情况都是U，除非主体权威性很高的，并…
5	觉着	U	T	一般情况都是U，除非是表达具体事实或者是…

Figure 1: 知识库示例

3.2 实验错误分析

本次评测中模型的错误多源于谓词叙实性在不同语境中的复杂多变。同一个谓词在人工语料与自然语料中的叙实性表现可能完全不同，例如“埋怨”虽属正叙实动词，但若说话人本身缺乏公正性，其后的命题真假便难以判定。因此语料中存在一些难以在简单的语句中进行判断的情况，需要结合对说话人的了解，更多的上下文语境，才能有效判断。这种情况难以通过专家知识准确界定具体的叙实性情况，也难以通过谓词替换等语义的方法来解决。亟需语言学家的进一步研究与探索。

4 结论

为提升大语言模型在叙实性推理任务中的表现，本文构建了一套结合RAG与谓词相似性的方法体系，并在DeepSeek-V3 上进行了验证，整体准确率达0.9240，其中人工语料为0.9615，自然语料为0.9099。系统通过四个模块（谓词分类、漂移条件获取、语境分析谓词替换、真值推理）有效应对叙实性漂移与多义词问题，显著提升了大模型在语义推理任务中的一致性与鲁棒性。结果表明，结合参数化与非参数化知识的推理机制是提升语言模型事实判断能力的有效路径。然而，对于依赖复杂语境或说话人背景的情况，模型仍存在一定误判，需进一步引入更丰富的语用知识与上下文理解机制。

第二十四届中国计算语言学大会论文集，第152页-第156页，济南，中国，2025年8月11日至14日。

参考文献

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, et al. 2024. Deepseek-v3 technical report.

Paul Kiparsky and Carol Kiparsky. 1970. Fact. In M. Bierwisch and K. Heidolph, editors, *Progress in Linguistics*, pages 143–147. Mouton, The Hague.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yu Wang and Yulin Yuan. 2021. A corpus-based study of factive verbs and its influencing factors. In *Workshop on Chinese Lexical Semantics*, pages 42–55. Springer.

李新良and 袁毓林. 2016. 反叙实动词宾语真假的语法条件及其概念动因. 当代语言学, (02):194–215.

李新良, 袁毓林, et al. 2023. 叙实性与事实性理论及其运用. 外语教学与研究出版社, 北京.

李新良. 2018. “感觉”类动词的叙实性及其漂移问题研究. 语言教学与研究, (05):65–75.

李新良. 2020. 现代汉语动词的叙实性研究. 北京大学出版社, 北京.

王昱and 袁毓林. 2020. 基于规则的双重否定识别——以“不v1 不v2” 为例. 中文信息学报, 36(4):12–19.

王昱. 2024. 双重否定结构自动识别研究. 中文信息学报, 38(2):36–45.

袁毓林. 2014. 隐性否定动词的叙实性和极项允准功能. 语言科学, (06):575–586.

袁毓林. 2020a. “忘记”类动词的叙实性漂移及其概念结构基础. 中国语文, (05):515–526, 638.

袁毓林. 2020b. “记得”的叙实性漂移及其概念结构基础. 语言教学与研究, (01):36–47.

袁毓林. 2020c. 叙实性和事实性：语言推理的两种导航机制. 语文研究, (01):1–9.