【LLM论文阅读】Google Research：REGEN数据集 && CF与LLM融合框架

原创 方方 方方的算法花园 2024年11月01日 08:44 北京

0 论文概况

1. 论文名称：
Beyond Retrieval: Generating Narratives in Conversational RecommenderSystems 《超越检索：在对话推荐系统中生成叙述》
2. 论文链接：https://arxiv.org/pdf/2410.16780
3. 论文作者所在机构：Google Research
4. 一句话概括：介绍了用于对话推荐系统的新数据集 REGEN 及其融合架构，阐述了相关工作、问题构建与基准测试、实验及分析等内容，旨在推动自然语言生成在推荐任务中的应用和发展。

1 论文出发点

本文的出发点是解决将推荐系统知识有效整合到大型语言模型（LLMs）中以实现自然语言生成（NLG）的挑战，特别是针对推荐任务的自然语言生成。
具体而言，主要基于以下背景和需求：
（1）对话推荐系统的发展需求：随着网络体验向更丰富和交互性方向发展，对话推荐系统成为前沿领域。此类系统利用自然语言生成，以吸引人的叙述形式提供个性化推荐，从而增强用户互动和满意度。然而，实现这一愿景面临着如何将生成语言与推荐及用户偏好相关联的挑战，需要有效整合用户和推荐系统内部状态的潜在表示与 LLM。
（2）现有数据集的不足：许多现有的数据集存在局限性，如集中于下一项推荐、结构化输出或短摘要，缺乏用于有效训练和评估对话推荐系统所需的丰富多样的对话元素。这促使作者创建一个更适合的数据集，以推动该领域的研究进展。
（3）改进推荐系统与语言模型融合方式的需求：尽管已有多种方法尝试将 LLMs 与推荐系统集成，但仍存在各种问题。例如，检索增强生成（RAG）方法的有效性依赖于检索模型准确性且存在计算开销；完全基于语言的技术面临信息编码和可扩展性问题；embedding输入型语言模型在生成能力和处理复杂叙事方面有待拓展。作者旨在提出一种新的融合架构，以更好地整合推荐系统知识，生成高质量、个性化的自然语言推荐。

2 论文贡献点

1.创建新数据集
生成了一个新的数据集 REGEN（Reviews Enhanced with GEnerative Narratives），该数据集包含了丰富的适合语言推荐任务的叙述内容，例如个性化的产品偏好解释、推荐商品的宣传以及用户购买历史的总结等。并且将其公开，以促进进一步的研究。
2.提出CF与LLM的融合架构
引入了一种融合架构（CF 模型与 LLM 的结合），这是针对 REGEN 数据集的基线。据作者所知，这是首次尝试分析 LLM 在理解推荐信号和生成丰富叙述方面的能力。实验表明，LLM 可以有效地从使用基于交互的 CF embedding的简单融合架构中学习，并且可以通过使用与

项目相关的元数据和个性化数据进一步增强。与单独使用任何一种类型的embedding相比，结合 CF 和内容embedding可使关键语言指标提高 4-12%。

3.建立基准和评估

使用知名的生成指标建立了基准，并使用评级 LLM 对新数据集进行了自动评估。

4.展示模型学习能力

通过多个示例表明模型学会了从历史中构建丰富的上下文和聚合叙述，而不是简单地记忆。

5.分析embedding贡献

分析了soft token embeddings，以确定模型在生成丰富叙述时如何结合内容和协作过滤信号。

3 REGEN 数据集

1. 数据来源：基于亚马逊产品评论数据集，选取“办公用品”和“服装、鞋类及珠宝”两个垂直领域作为数据基础，未来计划扩展到其他关键垂直领域及全部评论数据。

2. 数据集生成

（1）数据预处理：通过聚合用户评论并按时间戳排序创建用户序列，然后截断为最近的 50 项并过滤缺失标题的项目，得到办公用品和服装类别的相关数据统计。

（2）提示和叙述生成：旨在生成反映对话推荐系统中多样化交互的自然语言叙述数据集，通过设计不同维度（有无上下文信息、长短形式）的输出来理解其对生成语言质量和相关性的影响。利用 Gemini 1.5 Flash 模型生成数据，根据任务设计了相应的提示格式（如包含任务前缀、用户历史及元数据、预期输出格式等），并通过迭代评估提升生成质量。同时，利用 LLM 作为自动评估器评估生成输出的多个属性，如真实性、基础、清晰度、远见、个性化、用户丰富度和信心等，采用多阶段评估方法及集成评分过程确保评估的可靠性。

Generated output	***Description***
Product Endorsement	*A tailored product endorsement/sales pitch crafted based on understanding user's unique purchase history and reviews*
Purchase Reasons	*Concise explanations of the reasons behind a product recommendation.*
Purchase Reason Explanations	*Detailed and elaborate explanations of a purchase decision of the most recent item in context of the entire history.*
Brief User Summaries	*Concise summaries of a user's preferences and purchase history.*
Detailed User Summaries	*Comprehensive summaries of a user's preferences and history, potentially including specific examples and justifications.*
User Profiles	*A short phrase describing the type of user in product/shopping context.*

Table 2: Generated outputs and their descriptions. Average length is mean number of words for each output.

Rating Attribute	*Description*
Veracity	*Accuracy and truthfulness of the purchase reason, analyzing specific evidence and inconsistencies within the user's purchase history.*
Grounding	*Strength of the supporting evidence for all individual claims made in the purchase reason and explanation.*
Clarity & Specificity	*Clarity and specificity of the purchase reason and its explanation, and specificity of evidence.*
Foresight	*How much of the information used in the purchase reason came from the post-purchase review of the last purchased item.*
Personalization	*(User needs vs Product descriptions) Extent to which the narrative prioritizes information directly from the user's reviews and statements, as opposed to relying on product descriptions or general assumptions.*
User richness	*How information-rich the user's purchase and review history is for understanding their motivations and preferences.*
Confidence	*Confidence level in the evaluation, ranging from unreliable to high agreement between different evaluation methods.*

Table 3: Auto Rater: Rating Attributes.

3.评估方法

（1）总体方法：使用 Gemini Pro LLM 作为评估器，通过多次重复评分过程（10 - 15 次）并聚合分数得到最终得分，以提高评估结果的可靠性，并在数据集中标注用户的相应分数。

（2）评估指标和置信度评分：评估器根据七个属性（真实性、基础、清晰度与特异性、远见、个性化、用户丰富度和信心）对生成数据进行评分，分四个阶段进行评估（用户丰富度和信任度、购买原因和解释、用户总结、产品宣传），每个阶段生成相应的置信分数，通过多数规则和平均方法确定最终分数，以全面评估生成数据的质量。

4 CF与LLM的融合架构

1.任务定义

在用户交互序列背景下定义了对话推荐任务，将其分为三个关键任务：

（1）从近期上下文生成叙述，关注短期用户偏好；

（2）从聚合上下文生成叙述，旨在捕捉长期偏好；

（3）下一项预测，代表传统推荐目标。

论文主要聚焦于前两个任务。

2.模型架构

为有效整合推荐系统知识到 LLM 中进行自然语言生成，提出一种标准化架构，将协同过滤（CF）编码和用户交互的语义表示分离。

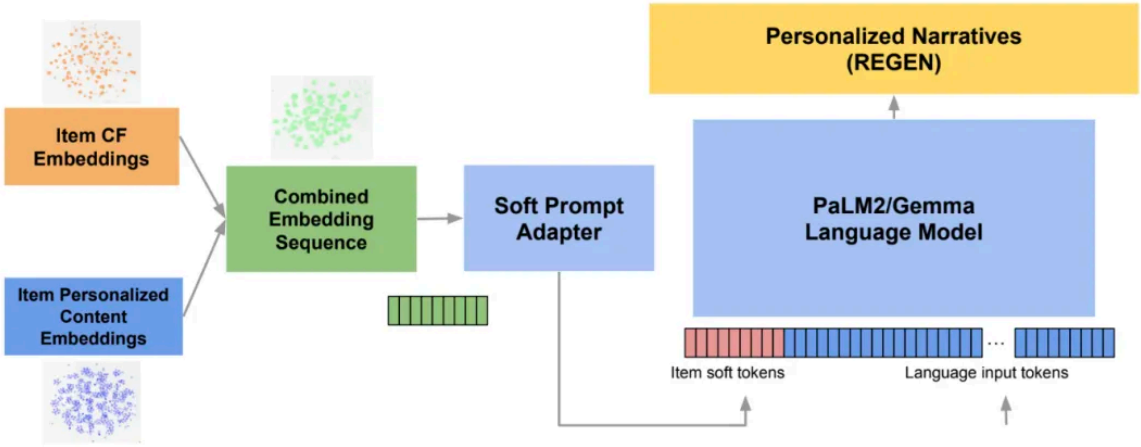

Figure 4: Overview of the Model Architecture.

- **CF Encoder**：处理用户 - 项目交互历史，通过构建用户 - 项目评分矩阵并应用加权交替最小二乘（WALS）矩阵分解获取用户和项目embedding。
- **Semantic Encoder**：语义编码器，捕获项目元数据和用户评论中的语义信息，通过连接两者并应用预训练句子embedding模型获得语义embedding。
- **Embedding Fusion**：embedding融合，对 CF 和语义embedding进行归一化处理后，通过简单拼接将两者融合，得到组合embedding，该预处理步骤可提升指标。
- **Adapter Model**：将组合 embedding 序列输入适配器模型（如 MLP 或 Transformer），生成与 LLM 模型维度匹配的soft token embeddings，作为用户交互历史的浓缩表示，以便与 LLM 集成。
- **Large Language Model (LLM)**：将适配器模型生成的soft token embeddings前置到文本prompt token embedding中，作为 LLM 的输入，使 LLM 能基于用户交互历史和prompt生成自然语言叙述。

5 实验验证

1. 数据集和任务设置

（1）数据集：使用 REGEN 数据集生成基准，包括 i）办公用品：包含 27K 个项目，以及 ii）服装、鞋子和珠宝：包含 376K 个项目，以评估模型在大小项目词汇表上的性能。

（2）任务：

- **叙述生成**：模型将用户的评分和评论历史作为输入，并生成一个综合叙述，总结他们的偏好和经历。对于语言输入，使用固定的prompt，本质上是提示模型根据观察到的历史生成相应的输出特征。
- **使用不同用户进行评估**：训练集和测试集包含不同的用户。迫使模型学习如何从提供的embedding中解读用户历史，而不是记忆特定用户。这确保了模型能够推广到新用户，并通过生成的叙述准确捕捉他们的偏好。

2. 评估指标

（1）传统自然语言处理指标： BLEU 和 ROUGE 得分来确立基准，提供性能趋势的高层次视图。

（2）相似性得分：使用基于句子embedding模型（如 Gecko）的相似性度量，将生成的叙述与 REGEN 参考叙述进行比较。这有助于评估与真实情况的语义接近程度。

（3）与评分者的并排评估：为生成叙述的质量和相关性提供更细致的评估，可以使用人工评分者或大模型评分。

3. 关键结论

（1）具有上下文的自然语言任务：语言输出基于即时上下文，例如，解释最近购买物品背后的推理和解释。与单独使用这些embedding中的最佳结果相比，将协同过滤（CF）和语义嵌入相结合，BLEU、ROUGE 和语义相似性指标分别提高了高达 12%、8% 和 8%。

（2）不依赖于上下文的自然语言任务：对于不依赖于即时上下文的任务（例如，根据综合用户档案生成叙述），将协同过滤和内容embedding相结合，与单独使用内容embedding相比，没有明显优势。可能是因为模型不需要特别关注最后一项物品，这在推荐任务中更常见，并且内容特征就足够了。

Output Feature	Metrics	CF	Content	CF + Content	(%) vs CF	(%) vs Content
Purchase reason	BLEU	17.71	21.07	21.92	+23.8%	+4.0%
	ROUGE-LSum	38.26	42.63	43.01	+12.4%	+0.9%
	Similarity	0.614	0.664	0.669	+9.0%	+0.8%
Purchase reason explanation	BLEU	19.56	21.68	24.24	+23.9%	11.8%
	ROUGE-LSum	37.26	38.05	41.11	+10.3%	+8.0%
	Similarity	0.783	0.787	0.851	+8.7%	8.1%
User Summary (Brief)	BLEU	17.96	19.3	19.12	+6.5%	-0.9%
	ROUGE-LSum	36.5	38.06	38.16	+4.5%	+0.3%
	Similarity	0.861	0.866	0.864	+0.3%	-0.2%
User Summary (Long)	BLEU	16.96	18.07	18.09	+6.7%	+0.1%
	ROUGE-LSum	31.61	33.35	33.42	+5.7%	+0.2%
	Similarity	0.885	0.895	0.903	+2.0%	+0.9%

Table 1: Performance benchmarks with using different embedding inputs to the PaLM2 XXS LLM. REGEN Office Products ($|I| = 27K$)

6 Embedding 分析

（1）Soft Prompt Embeddings与LM TokenEmbeddings 的关系

采用线性投影层作为适配器，将item embedding和content embedding连接后进行投影，得到soft prompt embeddings。通过 t-SNE 可视化发现，soft prompt及其组成部分（CF和content embedding）占据了一个与LM token空间不同的空间，这可能是由于LM Embeddings 空间的稀疏性导致的。尽管在当前实验条件下，适配器学习到了一个单独的embedding空间，但对于在大规模数据集上预训练的推荐语言模型，这种行为是否会持续仍是一个有待研究的问题。这一发现也与将item视为新（离散）tokens的概念相符，有效地扩展了语言模型的词汇表，但区别在于利用了完整的无约束embedding空间表示。

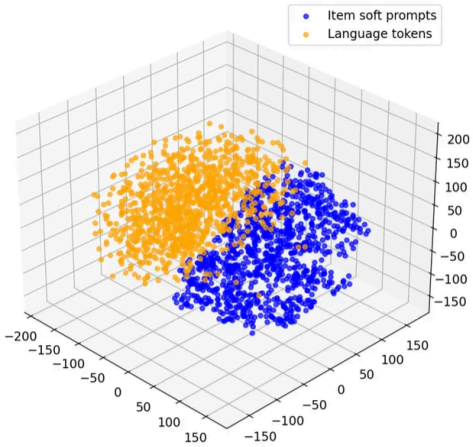

Figure 5: Comparing Soft prompt embeddings vs LLM token embeddings. Soft prompt embeddings generated for all items in Office Products.

（2）CF 与Content Embeddings的贡献

分析了投影后的 CF 和Content Embeddings在soft prompt embeddings总范数中的相对贡献分布，发现内容信号在学习到的soft tokens中占比更大，尽管 CF 部分也有一定贡献。这解释了实验中Content Embeddings模型优于 CF 嵌入模型，而两者结合效果最佳的现

象。以 Office Products 数据集为例，其中大部分项目仅出现在 10 - 20 个用户中，这导致 CF 信号有限。作者认为在具有更丰富交互数据的不同数据集上重复这些实验将有助于进一步了解 CF 和Content Embeddings的作用。

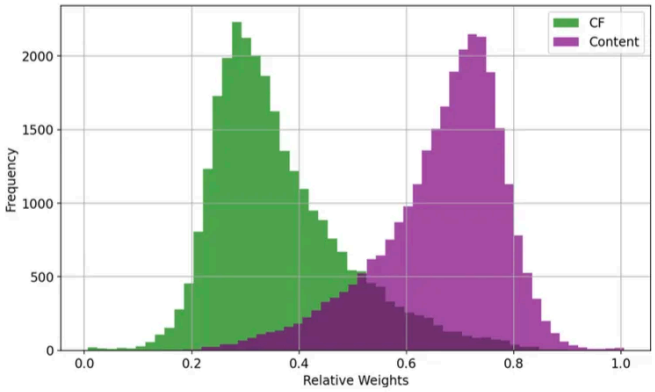

Figure 6: Comparing CF and Content contributions to Soft Prompts (Office Products).

7 附录

数据生成的prompt

购买原因和用户摘要使用few shot prompt，产品宣传使用zero shot prompt。

Prompts

购买原因和用户摘要的prompt（翻译版）

```
你是亚马逊卓越的用户你是亚马逊公司一位出色的客户互动专家。你的职责是了解客
你会收到用户一系列产品购买记录及其相应的评论。每个序列按时间顺序排列，从最

你的任务有两个方面。
- 其一，专注于他们购买序列中的最后一个产品（标记为 “Final Item: True”）
- 其二，利用购买和评论历史中的所有项目来识别该客户的购买属性。

以 JSON 格式提供以下信息：
1. **purchase reason**：从评论、产品描述和用户购买历史中可以推断出最后
2. **purchase reason explanation**：简要说明你确定购买原因的理由，考
3. **brief user summary**：根据用户的购买历史和评论，基于特定购物兴趣
4. **long user summary**：根据用户的购买历史和评论，基于特定购物兴趣提
5. **user profile**：根据用户过去的购买和评论，用 4 到 6 个词提供一个

示例JSON输出：
json
{
"purchase reason": ".......",
"purchase reason explanation": ".......",
"brief user summary": ".......",
"long user summary": ".......",
"user profile": "......."
```

```
}

**重要**
**请勿** 在你的答案中逐字引用用户评论的任何部分。
专注于见解，而非直接重复。
具体并与每个客户的购买和评论历史的背景相关。

以下是 7 位评论者的示例以及每个评论者 ID 序列中最后一项的预期答案。

reviewer id A01003458IEUPS8LQ1QU
Sequence Item Position 1
Final Item False
item title 罗乐德斯旋转开放式名片夹
item desc 独特的堆叠锁定功能使堆叠的托盘保持在一起，并具有锁定的稳定性
item price 9.989999771118164
overall 5
review title 五星
review text 这些托盘很好用，可以用来整理东西。我们为了工作而购买它们，它
"..."

"..."
Sequence Item Position 6
Final Item True
评估以下产品的购买意向。
item title 艾利双口袋文件夹，绿色，25 个装（47977）
item desc 用于打孔纸张的紧固件，用于散页的口袋——涵盖所有！双重用途使其非
item price 17.219999313354492
overall 5
review title 五星
review text 很棒的演示文件夹，我会再次购买它们。
End of Sequence

预期答案
{
"purchase reason": "需要用来存放销售演示材料的文件夹，这些材料包括散页
"purchase reason explanation": "客户需要一个能够存储由散页和打孔页组成
"brief user summary": ".......",
"long user summary": ".......",
"user profile":"......."
}
请（1）分析上次购买产品的购买原因（2）提供用户摘要，给出客户的产品购买和评
Current ID:
<current ID>
Customer Item/Review: <current reviews>
Answer:
```

Prompts

产品宣传的prompt（翻译版）

```
你是亚马逊卓越的用户你是亚马逊公司一位出色的客户互动专家。你的职责是了解
你会收到用户一系列产品购买记录及其相应的评论。
**每个序列按时间顺序排列，从最早到最近购买，由“Sequence Item Position”
您的任务是关注他们购买序列中的**last product**（标记为“Final Item: Tr

以JSON格式提供以下信息：
**product endorsement**：使用产品描述、用户购买历史和评论历史为最后一

示例JSON输出
json
{
"product endorsement": "........"
}
重要：
不要在回答中逐字引用用户评论的任何部分。
专注于见解，而不是直接重复。
具体并与每个客户的购买和评论历史的上下文相关。

Current ID:
<current ID>
Customer Item/Review: <current reviews>
Answer:
```

Prompts

生成结果示例（翻译版）

```
Generated Output 1
“购买原因”：“客户想要一个时尚的、相对较小的有多个口袋便于整理的手提袋。”
“购买原因解释”：“客户之前购买了一件衣服，正在寻找一个与之匹配的时尚手提包。
“用户简要总结”：“这位客户是一位喜欢购买时尚且舒适的服装的女性。她也对手提
“用户详细总结”：“这位客户是一位对时尚和舒适的服装有品味的女性。她购买了凉
“用户画像”：“时尚实用的女性”
“产品推荐”：“这款迈克高仕贝德福德顶部拉链口袋手提袋是一个奢华且实用的选择，
```

自动评分的prompt

```
您是一位高技能的客户行为分析师AI。
您将获得用户购买历史记录，包括商品详细信息、评论文本。
您的任务是回答有关用户购买行为的问题。
— BEGIN USER PURCHASE HISTORY —
User Purchase:在YYYY-MM-DD上购买了以下商品

— BEGIN RESULTS —
* Name: ...
Category: . . .
Description: . . .
Price: . . .
— END RESULTS —
User Purchase Review: . . .
...
— END USER PURCHASE HISTORY —

评估标准：
您的回复将根据以下内容进行评估：
* Accuracy：您的判断与购买的真实原因是否一致（如果有）
* Evidence：从评论历史中引用的示例的质量和相关性
* Reasoning：分析的清晰性和逻辑性
* Confidence：你的置信度如何反映证据的强度

Instructions:
* 仔细分析用户的评论历史，以识别偏好、需求或不喜欢的模式。
* 将建议的原因和解释与这些模式进行比较。
* 考虑积极和消极的证据。
* 明确引用评论历史中的具体例子来支持你的评估。
* 如果没有足够的证据做出强有力的判断，明确说明并解释原因。
* **对任何暗示对用户使用产品的体验有未来知识的购买原因进行严格惩罚**，给予非常低

附加考虑因素：
* 密切关注用户评论中使用的语言，特别是形容词、副词以及对特定特征或场合的任何提及
* 分别评估所提供购买原因的每个元素。如果原因提到特定属性（如颜色、大小、材料），
* 在评估购买原因时，优先考虑用户陈述和评论，而不是产品描述。只有当用户陈述直接证
* 分析看似不相关的购买，以寻找与当前购买原因的潜在联系。例如，过去购买礼物可能揭
* 如果某一原因特定方面的证据有限或完全基于假设，则对整体判断表示较低的信心。

Question:
请根据用户的购买历史评估所提供的用户资料、用户简要摘要和用户详细摘要的真实性。
— BEGIN USER PROFILE —
...
— END USER PROFILE —
— BEGIN USER BRIEF SUMMARY —
```

```
...
— END USER BRIEF SUMMARY —
— BEGIN USER LONG SUMMARY —
...
— END USER LONG SUMMARY —

Output:
对提供的用户画像、用户简要总结和用户详细总结的真实性进行详细评估。你的评估应包括：
* 根据评论历史中的证据，明确判断这些总结和画像是否真实、虚假或不确定。
* 从评论历史中支持你判断的具体例子。
* 你可以使用用户的整个购买历史来验证总结中的说法。
* 记住，即使对特定属性的负面评价也可以表明用户在意该方面。
* 避免做出没有直接证据支持的宽泛概括或推论。
* 专注于与总结中的具体说法有明确联系的有力、直接证据。
* 分析总结与用户过去行为之间的任何潜在不一致或矛盾。
* 考虑包括潜在矛盾信息在内的全部证据范围。
* 承认存在的复杂性和不确定性（如果有）。
* 你评估的总体置信水平（高、中、低）。

记住，你的最终判断必须是以下之一：
- Likely Very False
- Likely False
- Likely Somewhat False
- Uncertain
- Likely Somewhat True
- Likely True
- Likely Very True

请在回答之前大声思考你将如何回答这个问题并提供你的理由。
Answer:
```

END

#LLM学习 12 LLM与推荐 15 推荐系统 4 LLM论文阅读 13

#LLM学习 · 目录

上一篇
【prompt 自动优化】DSPy：原理与实践全解析

下一篇
【LLM论文阅读】Google Research：个性化语言提示的User Embedding模型