RAG工程如何评测?

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年10月03日 10:20 上海

◈ 技术总结专栏 ◈

作者：喜欢卷卷的瓦力

瓦力算法学研所
我们是一个致力于分享人工智能、机器学习和数据科学方面理论与应用知识的公众号。我...
117篇原创内容

公众号

本篇主要讲RAG工程的评测方法。

本篇属于RAG系列，上一篇整理了RAG的基础，没看过的小伙伴也可以参考~本篇来继续介绍RAG工程如何评测。下面是一个快捷目录。

一、RAG评估方法

有两种方法评估RAG：**独立评估**和**端到端**

1. 独立评估

独立评估就是对检索模块和生成模型分布评估。

1）检索模块

评估RAG检索模块性能的指标主要用于衡量系统（如搜索引擎、推荐系统或信息检索系统），即根据查询评估有效性。

具体指标包括：命中率 (Hit Rate)、平均排名倒数 (MRR)、归一化折扣累积增益 (NDCG)、精确度 (Precision) 等。这块跟推荐系统的评价指标相同。

- **命中率 (Hit Rate)**

检索结果中用户实际检索的实体词或者关键词所占的比例。

- **平均排名倒数 (MRR)**

是用来衡量返回结果的排名质量。MRR考虑了用户第一次遇到相关检索的排名；

结果列表中，第一个结果匹配，分数为1，第二个匹配分数为0.5，第n个匹配分数为1/n，如果没有匹配的句子分数为0。最终的分数为所有得分之和，再求平均。

计算方法

对于每个查询，首先计算倒数排名（即第一个相关检索的排名的倒数），如果没有相关检索结果，则倒数排名为0。然后，计算所有查询的倒数排名的平均值。

公式表示为：

$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$

其中，Q 是查询的集合，rank_i 是第 i 个查询中第一个相关项目的排名。

示例：

假设有3个查询，相关项目的排名分别是1, 3, 和 2。那么MRR为：

$\mathrm{MRR} = \frac{1}{3}\left(\frac{1}{1}+\frac{1}{3}+\frac{1}{2}\right) = \frac{1}{3}\left(1+\frac{1}{3}+\frac{1}{2}\right) \approx 0.611$

- **归一化折扣累积增益 (NDCG)**

NDCG用于衡量**排名质量。**它**考虑了所有相关结果的排名，并根据排名对其赋予不同的权重（排名越靠前，权重越大）。**

计算方法

首先计算DCG（Discounted Cumulative Gain），然后将其标准化。

DCG的计算方法为：

$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{\log_2(i+1)}$

其中，rel_i 是排名为 i 的项目的相关性分数，p 是考虑的排名位置。

接着，计算理想情况下（即按相关性排序）的IDCG（Ideal DCG），最后，NDCG为DCG与IDCG的比值。

$\mathrm{NDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p}$

示例：

假设对于一个查询，相关性分数（0-3，3为最相关）和排名如下：

- 实际排名：3, 2, 3, 0, 1
- 理想排名：3, 3, 2, 1, 0

计算DCG和IDCG，然后计算NDCG。

2）端到端评估

RAG 对特定输入生成的最终响应进行评估，主要是模型生成的答案与输入查询的相关性和一致性。

- 对无标签的内容评估评价指标：答案的准确性、相关性和无害性
- 有标签的内容评估评价指标：准确率 (Accuracy) 和精确匹配 (EM)

准确率比较简单，主要具体讲一下精准匹配 (EM)。

精确匹配是指模型给出的答案与参考答案完全一致时的评价指标。

如果模型的答案与参考答案完全相同，则EM得分为1；否则为0。

计算公式：

EM = 1，如果答案与参考答案完全一致；

EM = 0，如果答案与参考答案不一致。

二、RAG 的关键指标和能力

三个关键指标：答案的准确性、答案的相关性和上下文的相关性。

四个关键能力：主要是看抗噪声能力、拒绝无效回答能力、信息综合能力和反事实稳健性。

三、RAG的评估框架

这里介绍的主要是RAGAS 和 ARES。

1. RAGAS

RAGAS 是一个基于简单手写提示的评估框架，通过这些提示全自动地衡量答案的准确性、 相关性和上下文相关性。

算法原理：

1） 答案忠实度评估：利用大语言模型 (LLM) 分解答案为多个陈述，检验每个陈述与上下文 的一致性。即根据支持的陈述数量与总陈述数量的比例，计算出一个“忠实度得分”。

2） 答案相关性评估：使用大语言模型 (LLM) 创造可能的问题，并分析这些问题与原始问题的相似度。答案相关性得分是通过计算所有生成问题与原始问题相似度的平均值来得出的。

3）上下文相关性评估：运用大语言模型 (LLM) 筛选出直接与问题相关的句子，以这些句子占上下文总句子数量的比例来确定上下文相关性得分。

2. ARES

ARES 的目标是自动化评价 RAG 系统在上下文相关性、答案忠实度和答案相关性三个方面的性能。

ARES 减少了评估成本，通过使用少量的手动标注数据和合成数据，并应用预测驱动推理 (PDR) 提供统计置信区间，提高了评估的准确性。

算法原理：

1）生成合成数据集：ARES 首先使用语言模型从目标语料库中的文档生成合成问题和答案，创建正负两种样本。

2）训练大语言模型 (LLM) 裁判：然后，ARES 对轻量级语言模型进行微调，利用合成数据集训练它们以评其上下文相关性、答案忠实度和答案相关性。

3）基于置信区间对RAG系统排名：最后，ARES 使用这些裁判模型为 RAG 系统打分，并结合手动标注的验证集，采用 PPI 方法生成置信区间，从而可靠地评估RAG 系统的性能。

想要获取技术资料的同学欢迎关注公众号，进群一起交流~