

大作业组织形式

- 以小组为单位，每组不超过5人
- 以小组为单位进行评分，所有小组成员大作业成绩一样
- 实验报告首页应包含所有小组成员的个人信息

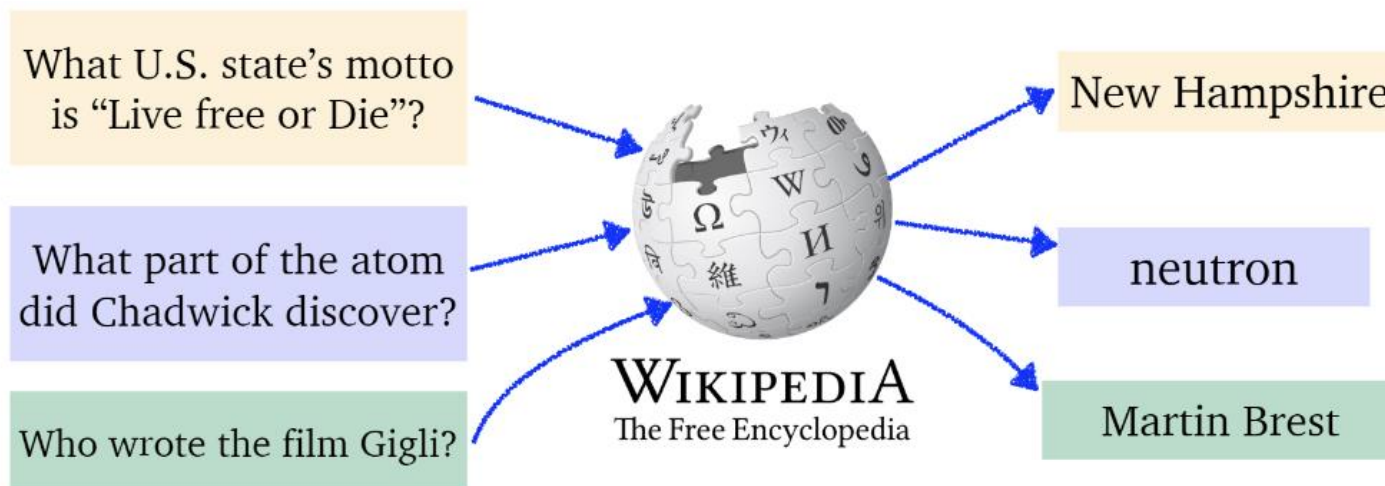
大作业内容

- 在WebQ数据集上进行检索
- 提交以下材料
 - a) 结果文件：检索结果hit@k
 - b) 程序代码：应具备对提交结果的检查功能
 - c) 实验报告：模型训练、验证、测试的过程

研究背景

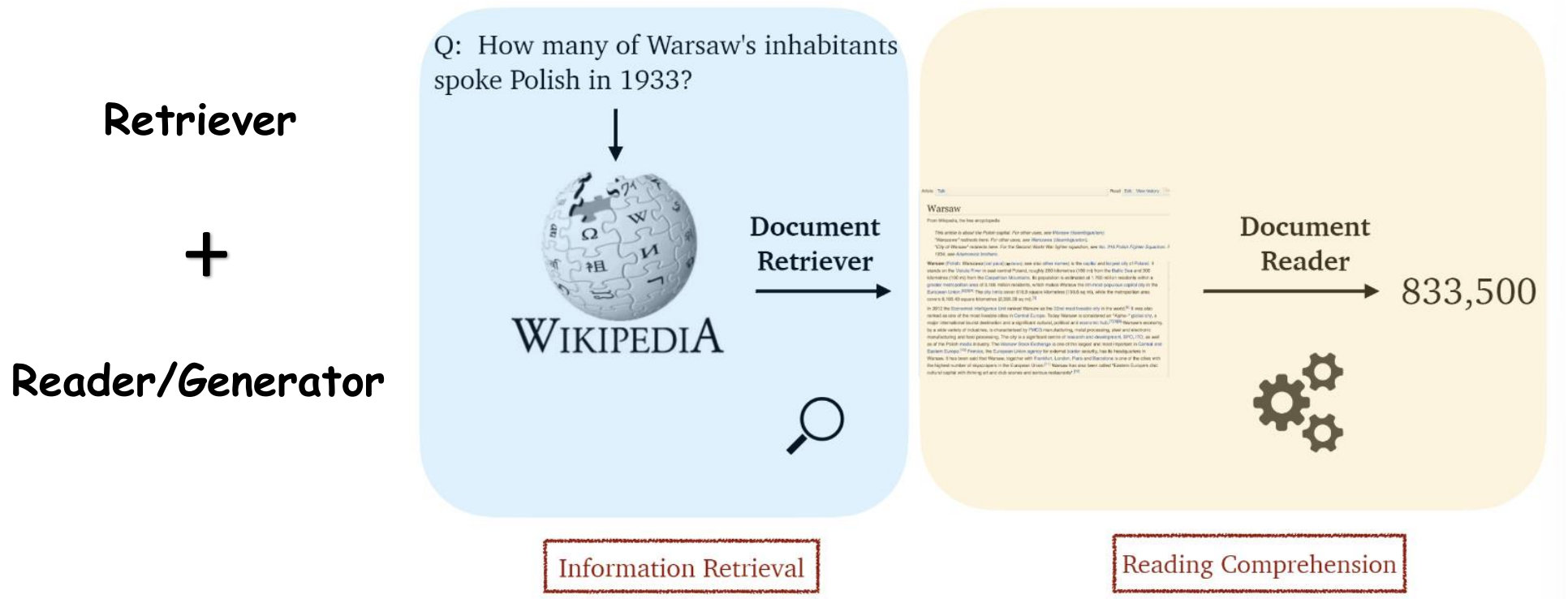
Problem setup: Open-domain Questions

- **Input:** question Q , English Wikipedia(~5 million documents) D
- **Output:** answer A



研究背景

Prevalent Architecture



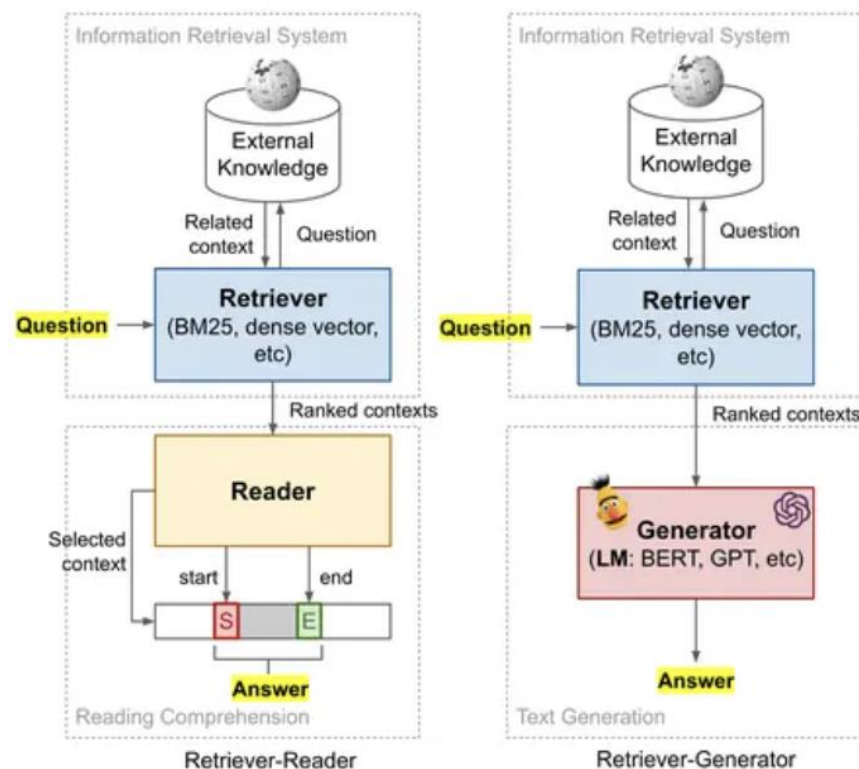
研究背景

开放域问答：极具挑战的任务

Retriever: 从大量上下文文档中根据 question 检索出包含答案相关的 documents 作为参考文档

Reader/Generator: 根据召回的相关文档提取或生成出问题答案

基于当前流行的架构，提高Retriever的召回准确率至关重要



检索任务

- 任务介绍
 - 要求：匹配WebQ中的查询信息，从大规模语料中检索出相关Passage。
 - 评价指标：hit@k，前k个检索结果中是否命中正确答案，对于单个query取值1/0
- 数据集
 - 由于原始语料较大，本次大作业只需在抽样处理后的数据上进行实验。
 - 可从课程网站上下载：IR_2024_Project.zip。

提交结果文件格式

- 提交的结果文件要求具体如下：

例如：

id	hit
----	-----

1	1
---	---

2	0
---	---

...

100	1
-----	---

- 每一对<id, hit>表示对测试集中的查询返回的前k个检索结果包含正确答案

使用的系统

- 可以使用开源工具
- 利用开源工具API实现自己的功能
- 实验报告中应提及所使用的系统

检索竞赛评分规则

- 参与形式：小组完成
- 检索效果（20分）：
 - 在给定的训练样本上进行模型训练，在给定的验证查询上进行模型验证选择
 - 汇报在200个测试查询上结果
 - 不得在测试查询上进行训练！违者视为作弊！
- 实验报告（10分）：
 - 对代码和运行方法进行说明
 - 详细描述实验中采用的技术
 - 对于提出的新方法、新技术有得分奖励
 - 新检索模型
 - 对现有模型、方法的提高和修正

实验报告

- 实现方案、主要代码类以及运行方法的说明
- 使用了什么技术？基于什么原理？如有必要给出公式
- 描述详细实验步骤
 - 数据处理
 - 模型训练
 - 验证测试
 - 要求能看出没有在测试查询集上进行训练
- 汇报在WebQ测试集中200个查询上的测试结果
- 明确给出最终提交的在测试集上得到的hit@1, 10, 100

结果提交

- 将所有材料做成一个压缩包，Email至 chenxinran22@mails.ucas.ac.cn
- 邮件标题：IR大作业_[组长姓名]
 - 源代码
 - 可执行程序
 - 结果文件
 - 实验报告
 - 但不提交中间文件，避免附件过大
 - 提交时限：2024年12月31日24点

提交材料的要求

- 代码清晰明确
- 建议使用Linux，推荐Ubuntu环境
- 实验报告中应明确说明如何运行程序
 - 要求“一键式”运行得到报告中的结果
 - 报告中明确给出需运行的脚本命令
 - 运行一个脚本命令（如bash或python），完成模型训练、模型选择、模型测试等步骤，得到报告中的测试结果
 - 说明最终产生的结果文件存放的位置（要求和打包提交的结果文件一致）
- 如需安装额外的软件包，应明确给出安装命令(例如sudo apt-get install xxx, conda install xxx, pip install xxx)