RAG-Fusion for Open-Domain Question Answering

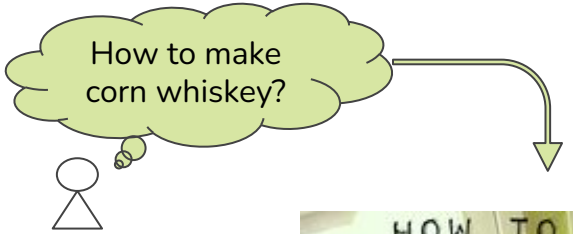
Xinao Han

Masterseminar
Suchmaschinen und Retrieval Augmented Generation
Sommersemester 2024

Introduction

Introduction

Introduction



We made a corn whiskey mash recently and documented the process for others to see...

Introduction

Query: corn mash

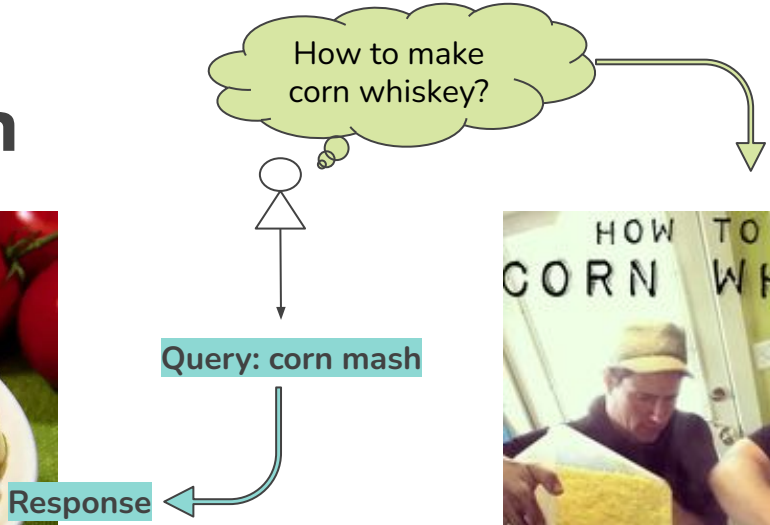
We made a corn whiskey mash recently and documented the process for others to see...

Introduction

Creamed corn is the perfect side dish to serve alongside comfort food favorites, such as meatloaf.

We made a corn whiskey mash recently and documented the process for others to see...

Introduction



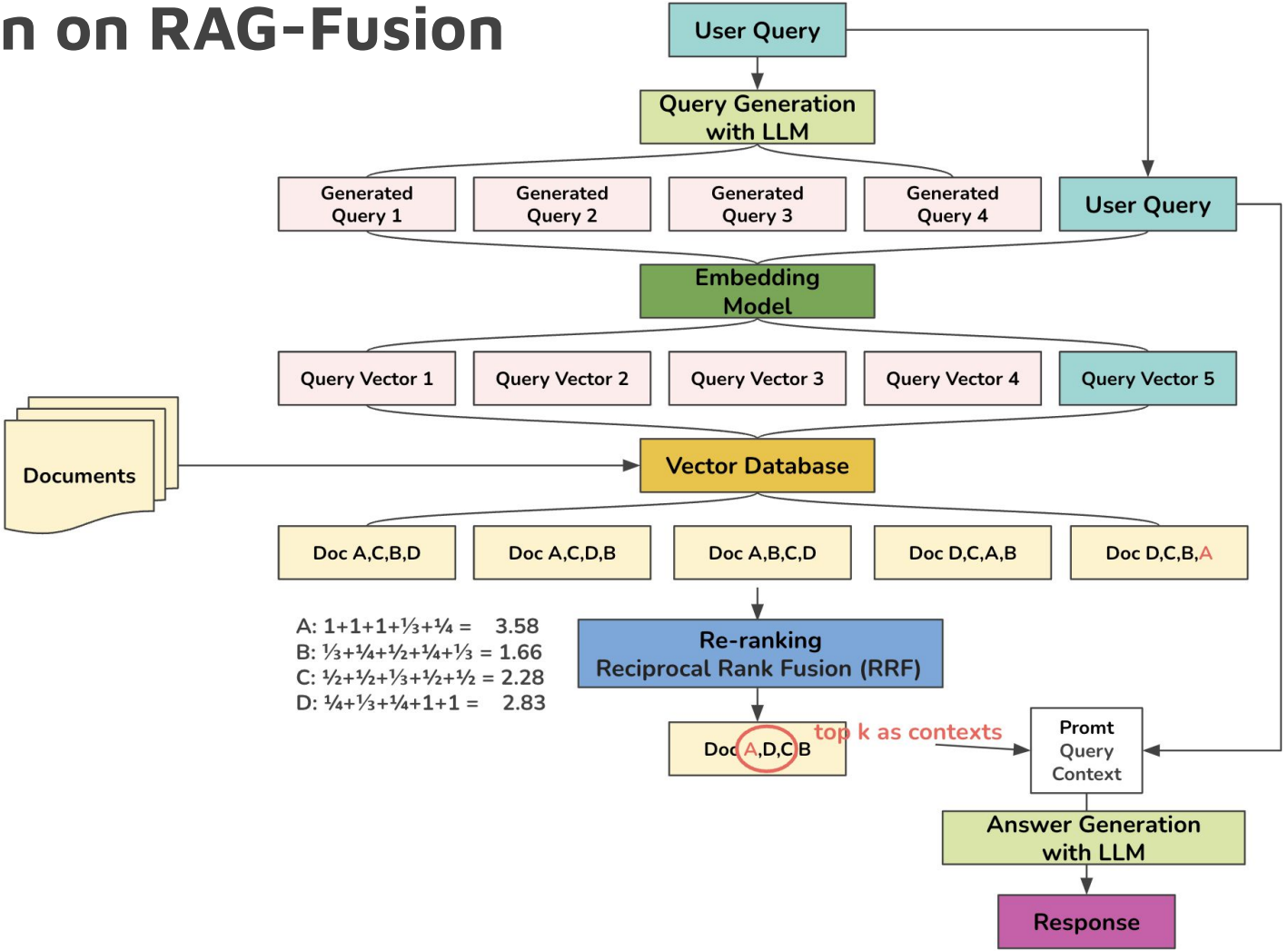
- **The problem of a naive RAG system**
  - Whenever we make a prompt, there is a difference between **WHAT WE ASK** and **WHAT WE INTEND TO ASK**.
  - " ... there is inevitably a **gap** between the input text and the needed knowledge in retrieval. (Ma et al., 2023)"
- **Naive RAG ⇒ Advanced RAG**

RAG-Fusion

Overview

- **What:**
  - It combines RAG and reciprocal rank fusion (RRF) by **generating multiple queries, reranking** them with **reciprocal scores** and fusing the documents and scores.
- **Who:**
  - Developed by ***Adrian Raudaschl*** *(*2023)
- **Why:**
  - To **bridge the gap** between traditional search paradigms and the multifaceted dimensions of human queries
- **Related Work:**
  - the **Infineon RAG-Fusion chatbot** for enhanced product information retrieval in Engineering and Account Management(Rackauckas, 2024)
- **In this work:**
  - Does RAG-Fusion also work well on open-domain QA task?

Zoom in on RAG-Fusion



Experimental Setup

Dataset

- MS MARCO V2 passage corpus

Query:

- 100 queries
- Official topics for the TREC Deep Learning (DL) 2023 shared task

Documents:

- Real word web documents
- Segmented into passages
- Each passage roughly contains between 500-1000 characters
- Total: 2M passages including ground truth for every query

Evaluation Metric

- **Information Retrieval**
  - mean reciprocal rank(MRR@10)

$$\text{MRR@10} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\min(\text{rank}_i, 10)}$$

- **Answer Generation**
  - Bleu, Rouge-1, Rouge-2, Rouge-L, Meteor

Implementation Details

- **RAG:**
  - LangChain
- **Embedding:**
  - "BAAI/bge-small-en-v1.5"
  - **B**AAI **g**eneral **e**mbedding (bge) model
  - Arch: bert
  - #para: 33.2M
- **Vector Store:**
  - DB Chroma
- **Generation:**
  - ChatGooglePalm
  - the Google **Pa**thway **L**anguage **M**odel(PaLM)
  - #Queries: 4

Results & Analysis

Information Retrieval

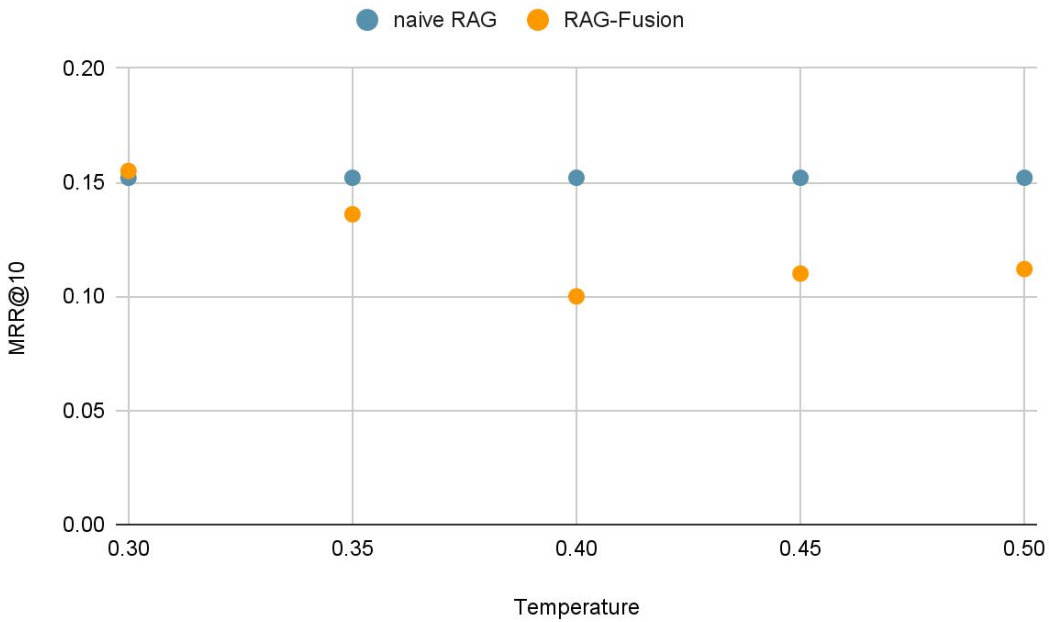


Figure 1: The impact of Temperature. The horizontal axis represents the value of Temperature and the vertical axis represents the MRR@10 metric.

- **Temperature**:
  - [0.0,1.0]
  - supervising - less surprising
- **Analysis:**
  - RAG-Fusion's performance varies significantly with different temperature settings.
  - The optimal temperature is **0.30**, where RAG-Fusion slightly outperforms naive RAG. Higher temperature settings (0.35 and above) lead to a decline in performance.
  - temperature<0.30:
    - Unstable generation
    - ChatGooglePalmError: ChatResponse must have at least one candidate

Query Generation 1

- **Original Query: "corn mash"**
  - ⇒ rank 50
- **Generate Similar Queries(temperature=0.3):**
  - "output": ["1. How to make corn mash",

    "2. Corn mash recipe",

    "3. Corn mash for whiskey",

    "4. Corn mash for moonshine"

    ]
  - ⇒ rank 7

Query Generation 2

- **Original Query: "similarity principle psychology definition"**
  - ⇒ rank 6
- **Generate Similar Queries(temperature=1.0):**
  - "output": ["1. Similarity principle in psychology definition",

    "2. Similarity principle examples",

    "3. Similarity principle in marketing",

    "4. Similarity principle in advertising"

    ]
  - ⇒ rank 29

Answer Generation

	Bleu	Rouge-1	Rouge-2	Rouge-L	Meteor
Naive RAG	0.010	0.181	0.036	0.168	0.168
RAG-Fusion (temp=0.3)	0.007	0.180	0.031	0.163	0.168

- **Analysis**
  - The naive RAG tends to perform slightly better across most evaluation metrics (BLEU, ROUGE-1, ROUGE-L), indicating that RAG-Fusion does not notably improve answer generation performance in this context.

Conclusions

- ➔ RAG-Fusion does not show a notable improvement over naive RAG in Open domain QA tasks.

- ➔ The parameter **Temperature** affects the quality of query generation, thereby influencing the performance of RAG-Fusion.
  - ◆ Lower temperatures generally lead to better performance for RAG-Fusion, provided that enough queries are generated.

Limitations & Future Work

Limitations

- **Ground Truth Source**:
  - The ground truth was derived from BM25 instead of being manually annotated, which makes the evaluation results not that reliable.

- **Long Execution Time**
  - Due to a more complex call to the LLM with multiple queries and more documents.
  - One of the major issues with the RAG-Fusion technique.

Future Work

- **Increase Repetition**:
  - Black-box Nature of LLMs
  - Conducting RAG-Fusion multiple times across different runs to improving result reliability.

- **Domain-specific QA tasks:**
  - The suboptimal performance of RAG-Fusion in open domain QA tasks does not imply that the RAG-Fusion technique itself is not worth researching.
  - Future work can explore the performance of RAG-Fusion in domain-specific QA tasks.

References

Abbasiantaeb et al., “LLM-Based Retrieval and Generation Pipelines for TREC Interactive Knowledge Assistance Track (iKAT) 2023.”

Gao et al., “Retrieval-Augmented Generation for Large Language Models.”

Ma et al., “Query Rewriting for Retrieval-Augmented Large Language Models.”

Zhu et al., “Large Language Models for Information Retrieval.”

https://arxiv.org/pdf/2305.14283

https://arxiv.org/pdf/2312.10997

Rackauckas, Z. (2024). RAG-Fusion: a New Take on Retrieval-Augmented Generation. *ArXiv, abs/2402.03367*.

- Introduction
- RAG-Fusion
- Experimental Setup
- Results & Analysis
- Limitations & Future Work

...Vielen Dank!

Vielen Dank!