Retreival Augmented Generation

András Kornai

2024 October 16

What we plan on doing

- Build a testbed
- Build a full system
- Do something for Hungarian
- Understand the theory better

Readings

- Today: Church, Sun Yue, Vickers, Saba, Chandrasekar 2024: Emerging trends: a gentle introduction to RAG
- Still looking for takers: Zhao and Rios 2024, https://huggingface.co/learn/cookbook/en/rag_evaluation

Separate tasks: 1. Infrastructure

- Download Hungarian WP **DONE** (Martin Juhász)
- Parse out tables and infoboxes **unassigned**
- Create a big database with lots of tables **unassigned**
- Create NL descriptions of tables 'Ez a táblázat Magyarország miniszterelnökeit listázza, dátumokkal' **unassigned**
- Verify DB purity (make sure stuff is not in other WPs) **unassigned**
- Create join chains "X was minister of Y government, Z is son of X" . . . **unassigned**

Separate tasks: 2. LLM

- Find a range of pre-existing LLMs
- Compare them on standard (Hungarian) tasks
- Maybe train adaptation layer
- Maybe add chain-of-thought prompting
- Test how these work

Separate tasks: 3. Retrieval

- Find appropriate docs outside WP
- Segment WP
- Construct retreival system based on keywords – Lucene
- Construct retreival system based on triple storage – if Istvan Soltesz finishes his port of OpenIE for Hungarian
- Construct retreival system based on vectors – Milvus
- Maybe combine these three, evaluate on its own merits

By Oct 30: build and test a baseline system

- Trello has the individual responsibilities
- We will have a zoom only meeting on the 23rd at 4PM (usual zoom link)
- We will use HuBERT for generating the vectors for the snippets
- Watch slack, we may have to move off of trello (monetization)