Retreival Augmented Generation

András Kornai

2024 October 9

WHAT WE PLAN ON DOING

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

READINGS

- Today: Self-RAG:
  https://openreview.net/forum?id=hSyW5go0v8
- Looking for takers: Zhao and Rios 2024, Church et al 2024,
- Also https://huggingface.co/learn/cookbook/en/rag_evaluation

Separate tasks: 1. Infrastructure

- Download Hungarian WP **DONE** (Martin Juhász)
- Parse out tables and infoboxes **pending** (István Soltész)
- 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
- Construct retreival system based on triple storage
- Construct retreival system based on vectors
- Maybe combine these three, evaluate on its own merits

SEPARATE TASK 4. REDUCE HALLUCINATION BY EXTERNAL TOOLS

- Find appropriate tool for arithmetic
- Find tool for SAT solving
- Create external call library for each
- Create test examples that exercise these
- Maybe add some other task reducible to SAT like path finding?
- Maybe add some other class(es) of task(s)?
- **assigned to 'NLP team'**

COMBINE TASKS 1: BUILD AND TEST THE RAG

- Integrate LLM from task 2 with Retrieval from task 3
- Measure on test obtained from task 1
- Integrate LLL from task to with External tools from task 4
- Measure on test obtained from task 1
- NLP team: Integrate 2+3+4 ?
- If so, also measure results