Retreival Augmented Generation

András Kornai

2024 October 2

What we plan on doing

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

Readings

- NO TAKERS SO FAR!
- Entity extraction, temporal reasoning: Zhao and Rios 2024
- Build-howto: Church et al 2024
- Eval-howto: https://huggingface.co/learn/cookbook/en/rag_evaluation
- Self-RAG: https://openreview.net/forum?id=hSyW5go0v8
- (Classic IR precursors: Xu and Croft 1996)

Separate tasks: 1. Infrastructure

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

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 tasks: 4. 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)?

Combine tasks: 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
- ? Integrate 2+3+4 ?
- If we do, also measure results