A RAG Approach for Typological Database Completion

Jonathan Hus, Antonios Anastasopoulos

Agenda

- Motivation
- Approach
- Results
- Conclusion

Motivation

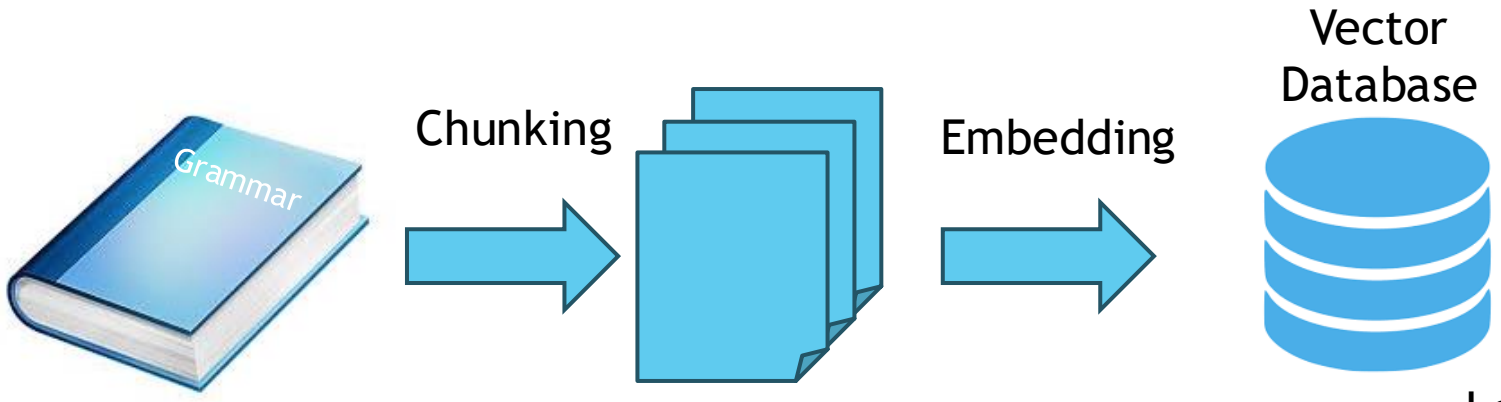
- Grambank is a massive online database that captures a wide range of grammatical phenomena in 195 features
  - Word order, verbal tense, nominal plurals, and other linguistic variables
  - Spans 215 different language families
  - Each feature has specific coding instructions. In addition to providing the feature code value, the human coders also provide the specific reference document (and ideally page number) that was used to inform the decision
- Grambank database currently covers 2467 languages
- Goal of Grambank is to cover all languages for which a grammar or grammar sketch exists

Can we use LLMs to assist with expanding coverage?

Approach

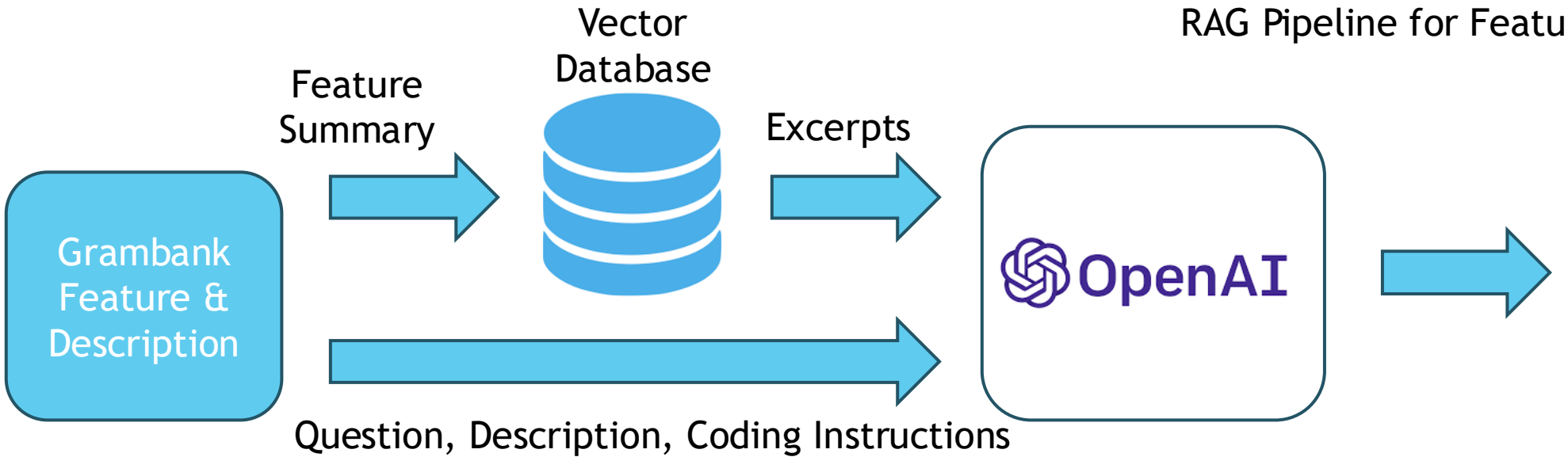
- Use LLMs and Retrieval Augmented Generation (RAG) techniques to identify language features
- RAG vector database is loaded with grammar books for languages
- To fill out a feature value for a specified language:
  - Relevant excerpts are extracted from the RAG database
  - The excerpts are provided to an LLM along with the feature description and coding instructions
  - LLM outputs the coding value for that feature

System Configuration



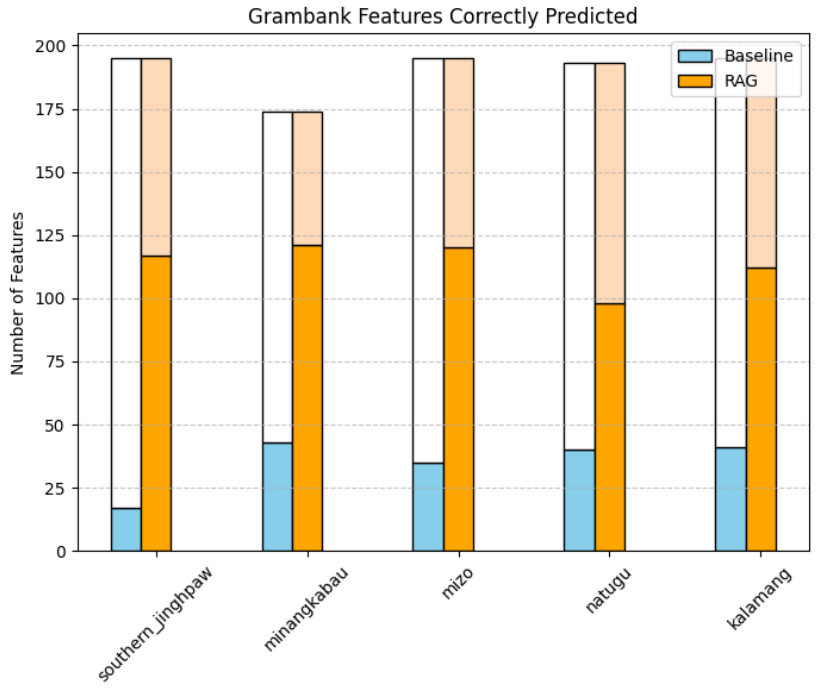
Loading DB with Reference Materials

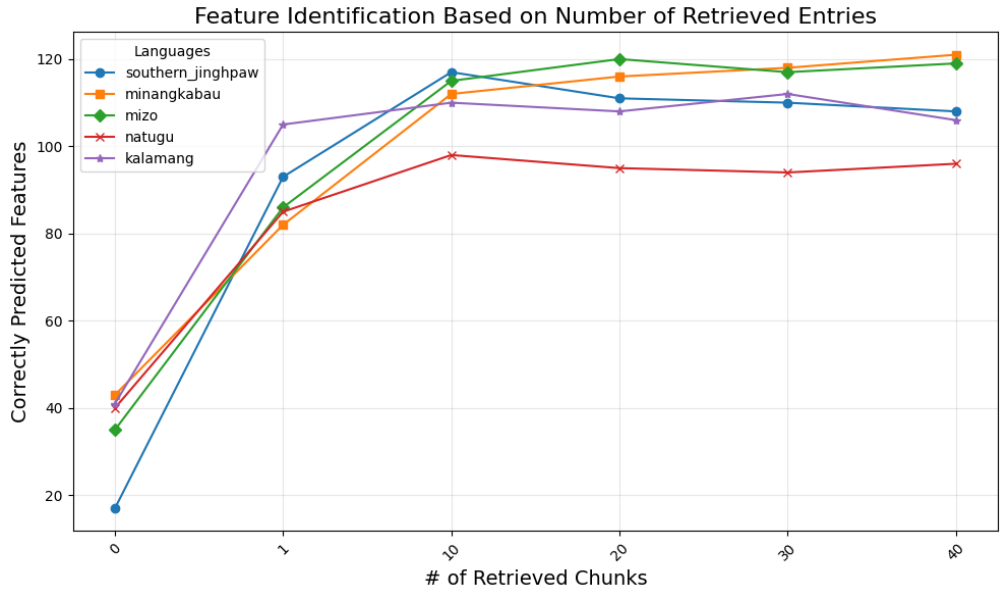
RAG Pipeline for Feature Completion



Baseline vs RAG Feature Prediction

- Compared performance when grammar book excerpts were included in the prompt vs not included
  - The 5 languages chosen are coded in Grambank so that “ground truth” values are available
  - Each language shows significant improvement in the number of features correctly predicted using RAG techniques
- Including more document excerpts in the prompt improves performance (up to a point)





Pipeline Analysis

- Vector database (RAG) and LLM can be analyzed separately to determine performance of each subsystem
  - RAG: Does it select the same grammar book pages as human coders?
  - LLM: Given the right excerpts from the grammars, does it properly code the feature?

RAG LLM

Language	Sources	Features w/ Pages	≥1 Match	Mean	Max	Skyline	Full Pipeline
Mizo (lus)						120/170	
– without summary	1	170	90	0.125	0.750		
– with summary	1	170	92	0.120	0.667		103/170
Southern Jinghpaw (kac)						112/195	
– without summary	3	195	134	0.143	0.571		
– with summary	3	195	132	0.144	0.500		117/195
Kalamang (kgv)						40/58	
– without summary	1	58	38	0.177	0.500		
– with summary	1	58	39	0.212	0.600		41/58
Minangkabau (min)						91/111	
– without summary	3	111	46	0.078	0.500		
– with summary	3	111	49	0.083	0.400		79/111
Natugu (ntu)						39/70	
– without summary	6	70	25	0.042	0.333		
– with summary	6	70	30	0.57	0.333		53/70

Conclusion

- Utilizing linguistic reference material such as grammar books into RAG pipelines can yield useful benefits
- Could potentially be useful as a tool for linguists completing Grambank entries
- Follow on work
  - Newer LLMs
  - Additional languages
  - Improve grammar book vectorization
  - Answer Verification


THANK YOU
GRACIAS
DANKE
ARIGATO
XIE XIE
GRACIAS
DANZE
MAHALO
MERCI
TAKK
TANKE
PANCIAS
DANKE
ARIGATO
XIE XIĚ
DANCI
KIITOS
DEHALO
GRACI
SUKRIA
DEEKUJE
DESEKUJE
ARIGATO
SZIĚKUJE
SHORAN
ARIGONE
DZIĚKUJE
DEIĚKUJO
MERCAI
PASBIUS
DERIKS
PHAUKES
EAMUUGE
MARRTAE
TESEKKÜR EDERIM
SBUN FAII LA
TORALI
EASONAS
DESCONAVO
AEEHAD