Expanding Horizons in RAG: Exploring and Extending the Limits of RAPTOR

Alex Laitenberger [1]

[1]Computer Science, Stanford

Problem

- LLM - lack on domain specific queries, document specific queries
- RAG - retrieval system using only contiguous chunks of text
- not capturing semantic relationships across document

Background: RAPTOR

Recursive Abstractive Processing for Tree-Organized Retrieval by Sarthi et al. (2024) - presented at ICLR:

- recursively clustering related text chunks and summarizing them
- hierarchical tree from the bottom up
- capturing both the meaning and the structural hierarchy

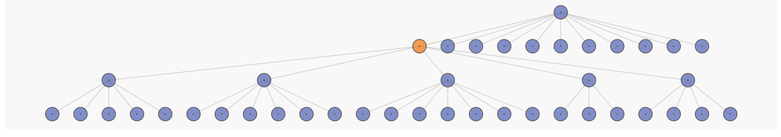

Figure 1. RAPTOR tree for Cinderella story.

Task: Analyse and improve the creation of retrieval trees in RAPTOR

Proposals :

Area	Suggestions
Alternatives to GMMs as base clustering algorithm	• Agglomerative • Neural
Chunking Text, Embeddings	• NLTK sentence tokenizer • Integrate positional embeddings – absolute – relative (e.g. rotary)
Summarization and tree building	• Merge sentence embeddings using a model similar to SBERT • Add a single root node to the tree with a full document summary

Figure 2. Suggestions for Improving RAPTOR Clustering Algorithm

Methods

Agglomerative Clustering - hierarchical - inherent tree structure - cluster creation by tree cuts

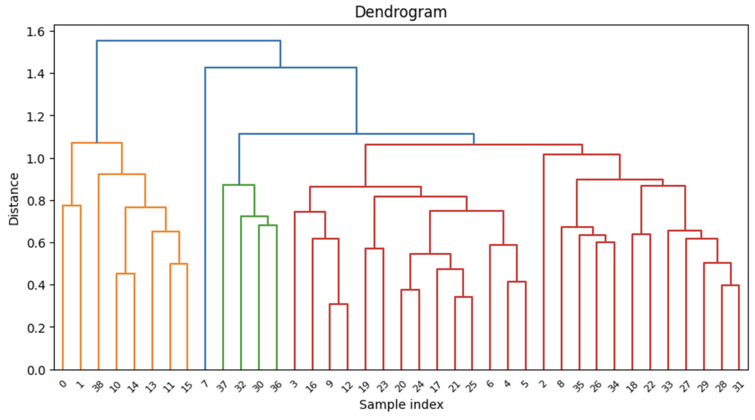

Figure 3. Dendrogram of Cinderella story using cosine distances and average linking.

Positional Embeddings - absolute positional embeddings

- 3 added elements: Startposition, distance to end, containment in section 1-3
- normalized and scaled to -0.1 to 0.1 to match SBERT embeddings
- added priority factor (5.0 for experiments)

Experiments

QASPER eval - 5,049 questions across 1,585 NLP papers

Accuracy in Answer F1 - General, Abstractive, Extractive, Boolean, None

Metric	Baseline	New tree	POS 5.0	POS 5.0 (II)
Answer F1	39.79%	38.07%	39.26%	36.01%
Answer F1 by Type				
Extractive	37.37%	39.17%	**40.19%**	37.53%
Abstractive	17.48%	18.86%	18.33%	**19.11%**
Boolean	**56.15%**	40.89%	42.62%	40.68%
None	**85.36%**	76.16%	80.88%	68.52%
Missing Predictions	64	31	34	42

Figure 4. QASPER Results for RAPTOR with SBERT and Gemini 1.0 Pro.

Analysis

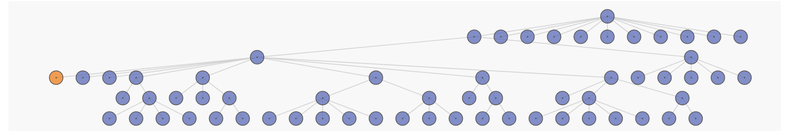

Figure 5. New RAPTOR tree for Cinderella story.

Qualitative answer comparison on Cinderella story - more nuances and more comprehensive response

Conclusion

Limitations of current RAPTOR in clustering algorithm

- Gaussian Mixture Models with dimensionality reduction
- flat trees - no applied soft clustering
- sentence tokenization

Agglomerative clustering

- tree-like clustering hierarchy displayable in dendrograms
- ideal base for building the originally suggested RAPTOR tree

Positional embeddings

- effect on clustering and tree buildling
- improvements on experiment results

Achievements

- full redesign of tree buildling - deeper balanced tree structures
- consistent outperforming on abstractive and extractive questions

Limitations

- QASPER results: decrease in none-type, boolean and strong deviations
- using QASPER F1-score as pure token-based comparisons

Future Avenues

- More experiments of different datasets (medical, legal, narrative)
- Smarter evaluations (model based judges)
- Explore relative positional embeddings, neural clustering, ways to realize soft-clustering

References

[1] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner.
A dataset of information-seeking questions and answers anchored in research papers, 2021.

[2] Patrick Lewis et al.
Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[3] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning.
Raptor: Recursive abstractive processing for tree-organized retrieval, 2024.

alaiten@stanford.edu