

문서 특징에 따른 RAG의 최적 청크 설정에 대한 연구

이금상¹, 이재환²

¹한림대학교 데이터사이언스학부 학부생

²한림대학교 경영학과 교수

dllrmatkd3254@naver.com, jaehwan@hallym.ac.kr

A Study on Optimal Parameter of Chunk in RAG based on Document Characteristics

Geumsang Lee¹, Jaehwan Lee²

¹Division of Data Science, Hallym University

²Dept. of Business Administration, Hallym University

요 약

RAG는 정보 검색과 생성 모델을 결합하여 주어진 주제나 질문에 관련된 지식을 생성하는 방법이다. 본 연구는 RAG의 성능을 높이기 위해 문서 내 문장의 평균 길이에 따른 청크의 크기와 오버랩 크기를 비교하여 최적화한다. 이를 통해 참조 문서의 특징에 맞춘 RAG를 개발할 수 있고, 다양한 종류의 글에 대해 맞춤형 답변을 제공할 수 있을 것으로 예상된다.

1. 서론

Large Language Model(LLM)은 자연어 처리 분야에서 제로샷 혹은 퓨샷에서 놀라운 일반화된 성능을 보여준다. 이런 주목할 만한 LLM의 성공에도 불구하고, 최신 정보나 특수 도메인 정보에 대한 답변에는 어려움을 겪고 있다. 이 어려움을 극복하기 위한 방법은 크게 두 가지가 있다. 첫 번째는 파인튜닝을 통해 모델의 파라미터를 직접 업데이트하는 방법이다. 하지만 이 방법은 데이터 구축 및 모델 학습 비용이 많이 든다는 단점이 있다[1][2].

두 번째는 Retrieval-Augmented Generation(RAG)이다. RAG는 의미론적 유사성을 계산을 통해 외부 지식에서 질문과 관련된 문서 청크를 검색함으로써 LLM의 환각을 완화할 수 있는 방법이다[3]. 이 방법은 모델을 직접 업데이트 할 필요가 없으며, 외부 지식을 참조함으로써 사실적으로 잘못된 콘텐츠를 생성하는 문제를 효과적으로 줄일 수 있다. 그러나 방대한 양의 외부지식 혹은 문서에 대해서 검색하기 위해서는 텍스트를 분할하는 작업이 필수적이다. 그리고 이러한 텍스트를 상황에 맞게 분할하기 위해서는 청크 사이즈와 오버랩 사이즈가 중요한 파라미터로 적용된다.

따라서 문서의 길이에 맞는 알맞은 청크 및 오버랩 사이즈는 검색 및 답변 성능에 중요한 영향을 미

친다. 본 연구에서는 여러 길이에 따른 문서별 청크 및 오버랩 사이즈를 분석하고 최적의 대안을 제안하고자 한다.

2. 문서 분할

주어진 문서에 대해 응용 프로그램 혹은 답변 생성 모델에 더 적합하도록 문서를 변환하고 싶은 경우는 빈번하다. 이를 위한 가장 간단한 방법은 긴 문서를 모델의 컨텍스트 창에 맞게 작은 사이즈로 분할 하는 방법이다. 하지만 긴 텍스트 문서를 여러 작은 덩어리로 분할하는 것도 잠재적인 복잡성이 많이 있다. 이상적으로는 의미상 관련된 텍스트 조각을 함께 유지하는 것이 좋다. 하지만 “의미상 관련”이 의미하는 바는 텍스트 유형에 따라 달라질 수 있다. 통상적으로 텍스트 분할기는 다음과 같이 작동한다.

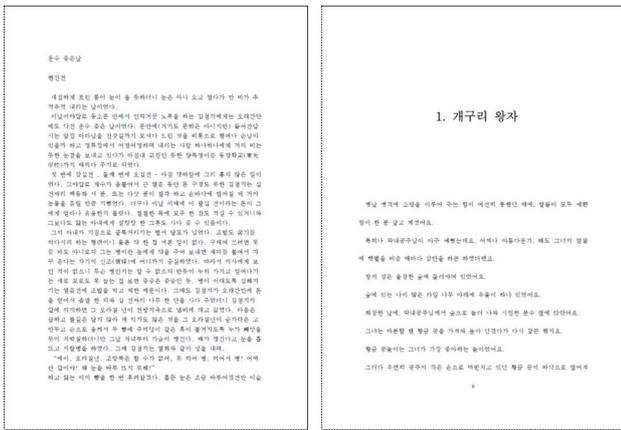
- (1) 텍스트를 의미상으로 의미 있는 작은 덩어리로 나눈다.
- (2) 특정 크기에 도달할 때까지 이러한 작은 덩어리를 더 큰 덩어리로 결합하기 시작한다.
- (3) 해당 크기에 도달하면 해당 청크를 고유한 텍스트 조각으로 만든 다음 내용이 약간 겹치게 작은 덩어리를 만든다.

여기서 긴 문서에서 분할된 작은 덩어리는 Chunk

를 의미한다. 또한 분할된 Chunk는 서로 고유한 텍스트 만을 가지는 것이 아니라 약간의 내용이 겹치게 Chunking 할 수 있다. 이를 Overlap 된 Chunk라고 정의한다. 따라서 본 연구는 최적의 Overlap 된 Chunk를 찾기 위해 노력한다.

3. 연구 방법

칭크 설정에 따른 RAG의 성능을 테스트하기 위해 현진건 작가의 ‘운수 좋은 날’ 문서와 아동 소설인 ‘개구리 왕자’ 두 문서를 수집하였다. 두 문서 모두 pdf의 형태로 구성되어있다. 그림 1과 같이 ‘운수 좋은 날’은 아동 소설 ‘개구리 왕자’에 비해 상대적으로 긴 문장과 단락으로 이루어져 있으며 아동 소설 ‘개구리 왕자’는 짧은 문장으로 이루어져 있다. ‘운수 좋은 날’은 10장으로 이루어져 있으며, ‘개구리 왕자’는 9장으로 이루어져 있는 문서이다.



(그림 1) 실험에 사용된 두 문서

4. 연구 결과

세팅 1: Chunk size = 400, Overlap size = 150

Query: 운수 좋은 날의 결말은 뭐야?

Answer

죄송합니다만, 제공된 정보에는 "운수 좋은 날"의 결말에 대한 정보가 포함되어 있지 않습니다.

세팅 2: Chunk size = 600, Overlap size = 50

Query: 운수 좋은 날의 결말은 뭐야?

Answer

"운수 좋은 날"의 결말은 김침지가 그의 아내가 죽은 것을 발견하고 그녀의 얼굴에 대고 중얼거리는 장면으로 끝납니다. 그는 아내에게 설령탕을 사다 놓았는데 왜 먹지를 못하냐고 묻습니다. 이는 김침지가 아내의 죽음을 받아들이지 못하고 있는 상황을

보여줍니다.

칭크 사이즈와 오버랩 크기에 따라 동일한 LLM도 다른 성능을 보이는 것을 확인할 수 있다. 향후 다양한 평가 query 셋을 구축하고, 여러 칭크 크기와 오버랩 크기에 대한 실험을 수행할 예정이다.

5. 결론

본 연구에서 여러 길이에 따른 문서별 칭크 및 오버랩 사이즈를 분석하고 최적의 대안을 제안하였다. 제안한 연구에 대한 성능은 문장과 문단의 길이의 차이가 있는 문서를 활용하여 테스트하였다. 제안된 연구에서는 다양한 글에 대한 최적의 칭크 및 오버랩 사이즈를 제안하며 각 문서에 최적의 질문 및 답변을 생성할 수 있는 성능을 보일 것이라 예상된다.

RAG를 구현할 때 문서의 특징에 따라서 Chunk 설정을 고려할 필요가 있다. 단, 본 연구 결과는 특정 도메인의 텍스트에서 수행되었으므로 연구 결과의 확장을 위해 다양한 도메인에서 반복 실험이 필요하다.

참고문헌

[1] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 100017.

[2] Gupta, A., Shirgaonkar, A., Balaguer, A. D. L., Silva, B., Holstein, D., Li, D., ... & Benara, V. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. *arXiv preprint arXiv:2401.08406*.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.