RAG における小説データベースの Chunk Size と Overlap Size と Embedding モデルの効果

阿部 晃弥 [1]　新納 浩幸 [2]
[1] 茨城大学工学部情報工学科　[2] 茨城大学大学院理工学研究科情報科学領域
20t4001l@vc.ibaraki.ac.jp　hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

大規模言語モデルにおいて，追加学習せずに外部知識を扱う手法として，RAG(Retrieval-Augmented Generation) が用いられる．これは，文書をベクトル化して保存しておき，正確な回答に必要な文書を prompt に組み込んで生成する手法である．本論文では，日本語の小説をデータベースとした RAG において，ベクトル・インデックスの作成の際，Chunk Size，Overlap Size 及び Embedding モデルを変更した場合の Retrieval の結果，回答への影響を検証した．実験の結果，Chunk Size には規則性が見られないが，Overlap Size は大きいほど概ね良い結果が確認できた．また，Chunk Size や Overlap Size の値とは無関係に，Embedding モデル間の性能差が確認された．

1　はじめに

近年，自然言語処理の分野では，ChatGPT を始めとする大規模言語モデル (Larage Language Model，以下 LLM と略す) が高い性能を示しており，多くの評価指標で従来のスコアを大きく上回る結果を残している．しかし，ある特定の分野の詳しい知識や比較的新しい知識といった学習データに含まれていない情報を扱うことが難しいという問題がある．LLM はその特性上，追加学習が困難であるという性質をもっており，その欠点を補うために，外部知識を様々な形で LLM に組み込むといった研究もおこなわれている．その一つが RAG(Retrieval-Augmented Generation)[1] である．RAG は，外部知識をベクトル・インデックスという形で保存し，入力文をクエリとしてインデックスに問い合わせ，関連する知識を入力文と合わせて LLM に与えることで，外部知識に基づいた生成を行わせることを目的とした手法である．一般的に，外部知識をテキストとして与える場合，テキストを Chunk Size ごとに切り分け，Embedding モデルを用いてベクトル化し，インデックスに格納する形式が用いられる．

本実験では，日本語の小説の内容を外部知識として与えた際，Chunk Size，Chunk 間の重複を認める Overlap Size が Retrieval の結果，回答へ与える影響について調査を行う．また，Chunk のベクトル化の際には，3 種類の Embedding モデルを利用する．それぞれのモデルで同様の条件で実験を行った場合の結果を比較し，その特徴や生成結果に関して考察を行った．

2　関連研究

RAG の基本的な考え方は LLM から得られない知識をデータベースから検索して得る形である．このため RAG の研究は主に検索手法を扱ったものが多い．HyDE[2] はクエリから関連文書を検索するのではなく，クエリの回答に役立つ関連文書を LLM から生成させ，その関連文書と類似の文書をデータベースから検索する．またクエリを小型の LLM を利用することで，RAG に有効な形に書き換える研究もある [3]．ただし RAG の様々な設定はクエリ依存の部分が多く，これらの手法が汎用的に有効であるとは言いがたい．本論文はこのような拡張した手法を提案するものではなく，基本的な RAG の検索部分の設定パラメータを調整することで，性能にどの程度の差異が生じるかを調べている．この点から本論文と最も関連が深いのは論文 [4] である．そこでは LLM の検索部の影響を系統的に調査している．

また RAG の検索部以外の改良としては Self-RAG[5] がある．RAG では外部知識を併用することで逆に LLM の性能が悪化する場合もある．そこで Self-RAG では Reflection Token を導入し，検索された文書の関連のある/なし，支持の度合い，有益度などを判断することで，徐々に回答の生成文を作成してゆく．本論文ではクエリが（マイナーな）小説の

内容に関するものであり，LLM 内に回答の知識がないことを前提としている．

3 実験

QA タスク用の RAG を実装し，Chunk の作成時のパラメータ，Retrieval に用いる Embedding モデルを変更することによる Retrieval の結果，回答の生成に与える影響について調査する実験を行った．

3.1 実験用データセット

3.1.1 実験用データベース

青空文庫[1]に公開されている太宰治の小説「女生徒」,「千代女」本文を，句点で分割してデータベースを作成した．なお，作品名と作者名を識別するため，一文毎に文頭に「小説名（作者名）：」を付与した．

3.1.2 実験用テストデータ

RAG の評価を行うために，質問，対応する本文，正答からなる評価用のデータセットを作成した．この質問は，太宰治の小説「女生徒」の内容に関するものであり，正答は，質問の正解となる．また，対応する本文は，それぞれの質問に正確に回答するために必要となる本文の文章となる．この質問，対応する本文，正答の組を 15 組用意し，実験のテスト，評価に用いた．本テストデータは付録 A に記した．

3.2 RAG の設定

本実験の RAG は LangChain を用いて実装した．実験の詳細な設定に関して，以下に示す．

3.2.1 Chunk の作成

Chunk の作成には，3.1.1 章の実験用データベースを用いた．Chunk Size は実験ごとに変更し，100 トークンから 1000 トークンまで 100 トークンずつの 10 種類である．Overlap Size も実験ごとに変更し，Chunk Size の 0%，25%，50%の 3 種類である．よって，作成する Chunk は 30 種類となる．また，Chunk の最小単位は小説の本文一文とする．

3.2.2 Index の作成

3.2.1 章で作成した Chunk を埋め込みモデルに与えて Index を作成した．埋め込みモデルには，HuggingFace で提供されている paraphrase-multilingual-mpnet-base-v2[2], intfloat の multilingual-e5-large[3]，OpenAI 社から提供されている API から text-embedding-ada-002[4] の 3 種類を用いる．

3.2.3 Retrieval

3.2.2 章で作成した Index を用いて Retrieval を行う．具体的には，質問に関連する Chunk を Index から取得し，回答の生成の際に prompt に埋め込んで使用する．今回の実験では，Index から取得する Chunk 数を，質問とのコサイン類似度上位 4 つとした．

3.2.4 Generation

回答の生成には，LINE 社が公開しているモデル japanese-large-lm-3.6b-instruction-sft[5] を使用した．

3.3 実験結果

実験用データセットで作成したデータベースを実装した RAG に与え，評価用データの 15 個の質問に回答させた．それぞれの質問には，4 つの Chunk が Retrieval され，回答生成に用いられた．そして，Retrieval された Chunk に実際に質問に関連する文章が含まれているか，また，回答が正確であるかを評価した．この実験を，各 Chunk Size，各 Overlap Size，各 Embedding モデルに応じて行った．

Retrieval 成功数，回答の成功数に関して，paraphrase-multilingual-mpnet-base-v2 の結果を表 1，intfloat/multilingual-e5-large の結果を表 2，text-embedding-ada-002 の結果を表 3 に示す．[6] 各モデルの Overlap Size ごとの Retrieval 成功数及び回答の成功数の平均値に注目すると，どのモデルに関しても Overlap Size が大きくなるほど，ほとんどの数値が増加している．よって，日本語の小説をデータベースとした RAG の場合，Overlap Size が大きくなるほど，より良い Retrieval，生成結果が得られる可能性があると考えられる．

1） https://www.aozora.gr.jp/

2） https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

3） https://huggingface.co/intfloat/multilingual-e5-large

4） https://openai.com/blog/new-and-improved-embedding-model

5） https://huggingface.co/line-corporation/japanese-large-lm-3.6b-instruction-sft

6） OvSz は Overlap Size，CkSz は Chunk Size，RetS は Retrieval の成功数，正答は質問に対する回答の正答数を示す．また，平均値は各 Overlap Size に対する Chunk Size 全体の Retrieval の成功数と回答の正答数の平均値を示す．

表 1 paraphrase-multilingual-mpnet-base-v2 の実験結果

OvSz	CkSz	RetS	正答	CkSz	RetS	正答
0%	100	4	2	600	3	1
	200	4	4	700	5	0
	300	6	2	800	2	1
	400	6	3	900	3	1
	500	3	1	1000	5	0
平均値					4.1	1.5
25%	100	4	2	600	3	1
	200	6	4	700	4	2
	300	7	3	800	6	3
	400	5	2	900	0	0
	500	4	2	1000	2	0
平均値					4.1	1.9
50%	100	4	1	600	6	2
	200	7	3	700	5	3
	300	7	4	800	6	1
	400	6	4	900	6	1
	500	6	2	1000	6	2
平均値					5.9	2.3

表 2 intfloat/multilingual-e5-large の実験結果

OvSz	CkSz	RetS	正答	CkSz	RetS	正答
0%	100	7	3	600	9	4
	200	9	5	700	9	3
	300	8	2	800	5	1
	400	7	3	900	10	6
	500	11	6	1000	6	0
平均値					8.1	3.3
25%	100	8	3	600	9	5
	200	9	3	700	8	1
	300	9	5	800	7	4
	400	10	6	900	6	5
	500	9	5	1000	11	1
平均値					8.6	3.8
50%	100	12	4	600	10	6
	200	7	4	700	10	4
	300	10	5	800	8	3
	400	7	4	900	8	3
	500	11	4	1000	8	0
平均値					9.1	3.7

また，Retrieval 成功数に関して，Overlap Size ごとの比較をグラフで表した．Overlap Size 0%を図 1，Overlap Size 25%を図 2，Overlap Size 50%を図 3 に示す．それぞれのグラフでは，Chunk Size の大きさという観点では規則性は見られず，多くの Chunk Size で緑色の Multilingual-e5-large が良い性能を示しており，赤色の paraphrase-multilingual-mpnet-base-v2 が劣った性能を示している．ここで，今回の実験は限定的な状況のため，一概に Multilingual-e5-large が優秀なモデルであり，paraphrase-multilingual-mpnet-base-v2 が劣ったモデルであるとはいえない．しかしながら，一つのタスクに限定すると，Chunk Size

表 3 text-embedding-ada-002 の実験結果

OvSz	CkSz	RetS	正答	CkSz	RetS	正答
0%	100	9	5	600	4	2
	200	5	3	700	4	2
	300	8	4	800	6	2
	400	7	2	900	7	2
	500	7	2	1000	7	0
平均値					6.4	2.4
25%	100	10	7	600	5	2
	200	8	5	700	6	1
	300	8	4	800	7	2
	400	5	3	900	4	0
	500	7	3	1000	3	0
平均値					6.3	2.7
50%	100	9	5	600	4	2
	200	7	4	700	6	4
	300	7	4	800	6	2
	400	9	4	900	7	5
	500	7	3	1000	8	0
平均値					7.0	3.3

や Overlap Size とは無関係に Embedding モデルの性能差があらわれることがわかった．

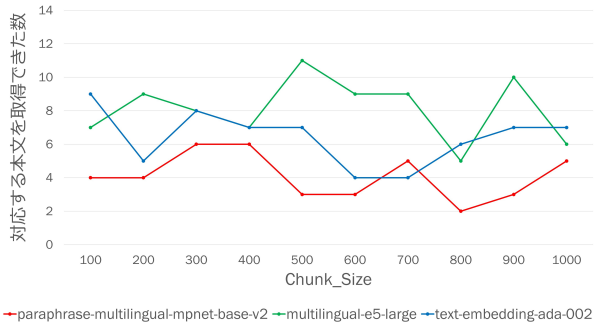

図 1 Overlap Size 0%の場合の各モデルの比較

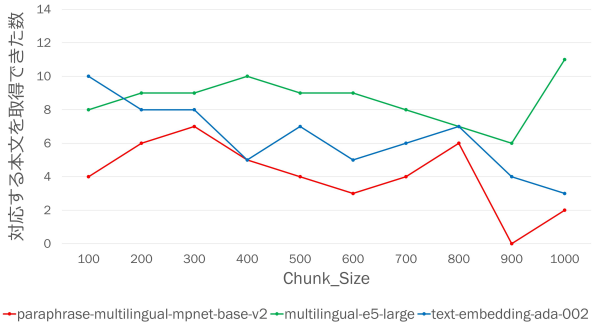

図 2 Overlap Size 25%の場合の各モデルの比較

4 考察

事前学習で得られなかった知識に関して，LLM が正確な回答を生成するには，正しい Retrieval が必要であると考えられる．つまり，RAG においては Retrieval の成功数と質問に対する回答の正答数に

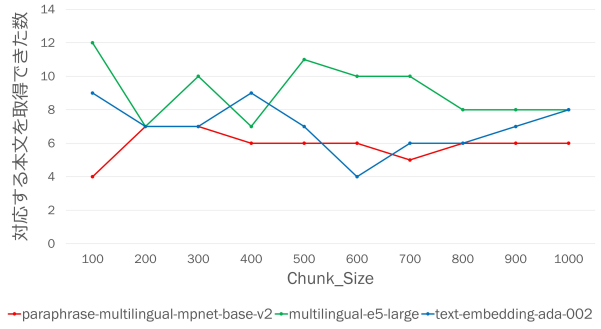

図 3　Overlap Size 50%の場合の各モデルの比較

は，相関関係があると考えられる．そこで，3.3 章で得られた実験結果について，Chunk Size，Overlap Size の組ごとの Retrieval の成功数と回答の正答数について，相関係数を調査し，さらに濃度分布図に表した．作成した濃度分布図を図 4 に示す．調査した結果，相関係数は **0.645** であり，Retrieval の成功数と回答の正答数について相関が示された．また，図 4 から，視覚的にも相関関係があることがうかがえる．

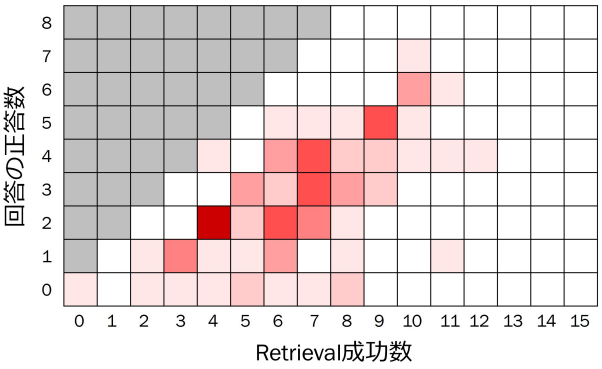

図 4　Retrieval 成功数と正答数の関係

次に，テストデータの質問ごとに各モデルの Retrieval 成功数を表にあらわした．その結果を表 4 に示す．[7] この表から，それぞれの質問についての傾向が読み取れる．

質問 1，2 に関しては，いずれのモデルであってもほとんど Retrieval できていない．その原因の一つとして，Retrieval したい情報の前後が全く異なる文脈であるため，情報が埋もれてしまっていることが考えられる．また，その他の原因として，似通った意味の情報が本文中に多く存在し，上位 4 つに Retrieval できていないということが考えられる．

そして，一部のモデルでは Retrieval が可能だが，

7）　QzNo. は質問番号を示す．mpnet, e5, ada は各 Embedding モデルを示し，それぞれ paraphrase-multilingual-mpnet-base-v2, intfloat/multilingual-e5-large，text-embedding-ada-002 である．

表 4　各質問の Retrieval 成功回数

QzNo.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
mpnet	0	0	3	12	8	9	17	22	22	20	20	1	0	8	0
e5	1	2	13	14	3	28	24	12	29	26	26	21	28	20	11
ada	3	2	11	22	1	21	24	18	24	14	13	2	13	18	11

一部のモデルではほとんど Retrieval できないというようにモデルによる得手不得手な質問が存在している．例えば，質問 5 は paraphrase-multilingual-mpnet-base-v2 では Retrieval できているが，他の 2 つのモデルでは Retrieval できていない．一方で質問 12 は，intfloat/multilingual-e5-large では Retrieval できているが，他の 2 つのモデルでは Retrieval できていない．

さらに，各 Chunk Size, 各 Overlap Size, 各 Embedding モデルでそれぞれの質問が Retrieval できたかどうかを図に示した．その一例として，質問 7 の Retrieval の結果を図 5 に示す．この図の赤く囲んだ部分に注目すると，Overlap Size 25%の場合，Chunk Size 600 から 900 にかけて不自然に Retrieval に失敗していることがわかる．よって，クエリに応じて情報が埋めにくい Chunk Size，Overlap Size が存在すると考えることができる．

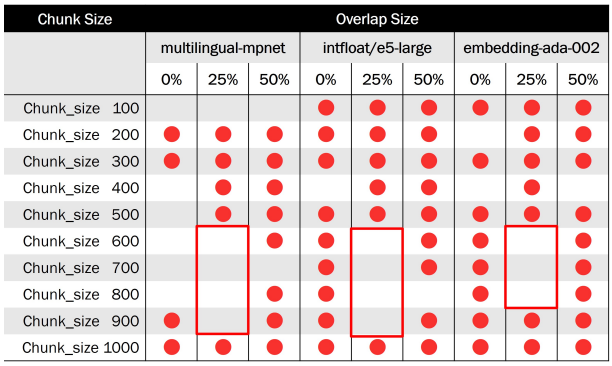

図 5　質問 7 の Retrieval の結果

5　おわりに

本研究では，LangChain から作成した RAG を用いて，日本語の小説として太宰治の作品をデータベースとして Chunk Size，Overlap Size，Embedding モデルをそれぞれ変更した際の影響について調査を行った．そして，Retrieval の成功数，小説の内容に関した質問に対する回答の正答数という観点から評価を行った．その結果，Chunk Size に決まった規則性は見られないこと，Overlap Size が大きいほど RAG の性能が向上すること，Embedding モデル間の性能の差は Overlap を変更しても変わらないことがわかった．

謝辞

本研究は国立国語研究所の共同研究プロジェクト「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」及び JSPS 科研費 23K11212 の助成を受けています.

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[2] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022.

[3] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 5303–5315, Singapore, December 2023. Association for Computational Linguistics.

[4] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation, 2023.

[5] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.

A　実験用テストデータ

3.1.2 章で述べた RAG の実験に用いた評価用のテストデータを以下の表 5 に示す.

表 5　テスト用の質問と対応する本文と正答

No.	質問の内容	対応する本文	正答
1	太宰治の小説である「女生徒」の主人公のいとこの名前は何ですか？	一通は、私へ、いとこの順二さんから。	順二
2	太宰治の小説である「女生徒」の主人公の一番好きな子の名前は何ですか？	私は、親類中で、いや、世界中で、一ばん新ちゃんを好きだ。	新ちゃん
3	太宰治の小説である「女生徒」の主人公の家に来たお客さんの名前は何ですか？	きょうのお客様は、ことにも憂うつ。大森の今井田さん御夫婦に、ことし七つの良夫さん。先刻、今井田が来ていたときに、お母さんを、こっそり恨んだことを、恥ずかしく思う。今井田さん、おかえりになる。	今井田さん御夫妻
4	太宰治の小説である「女生徒」に登場する美しい青色の似合う先生の名前は何ですか？	けさの小杉先生は綺麗。私の風呂敷みたいに綺麗。美しい青色の似合う先生。山中、湖畔の古城に住んでいる令嬢、そんな感じがある。	小杉先生
5	太宰治の小説である「女生徒」で描かれているのは何月何日の話ですか？	けさから五月、そう思うと、なんだか少し浮き浮きして来た。	5 月 1 日
6	太宰治の小説である「女生徒」で「私」が可愛がったペットの名前は何ですか？	ジャピイと、カア（可哀想な犬だから、カアと呼ぶんだ）と、二匹もつれ合いながら、走って来た。二匹をまえに並べて置いて、ジャピイだけを、うんと可愛がってやった。縁側に腰かけて、ジャピイの頭を撫でてやりながら、目に浸みる青葉を見ていると、情なくなって、土の上に坐りたいような気持になった。あんまり面白くて、この木をゆすぶって、ボタボタ落としたら、ジャピイ夢中になって食べはじめた。ばかなやつ。茱萸を食べる犬なんて、はじめてだ。急に、歯ぎしりするほどジャピイを可愛くなっちゃって、シッポを強く掴むと、ジャピイは私の手を柔かく噛んだ。	ジャピイ
7	太宰治の小説である「女生徒」で主人公が今年初めて食べたものは何ですか？	ことし、はじめて、キウリをたべる。	キウリ
8	太宰治の小説である「女生徒」でお昼御飯のときに出たお話は何ですか？	お昼御飯のときは、お化け話が出る。ヤスベエねえちゃんの、一高七不思議の一つ、「開かずの扉」には、もう、みんな、きゃあ、きゃあ。それからまた、ひとしきり恐怖物語にみなさん夢中。それから、これは怪談ではないけれど、「久原房之助」の話、おかしい、おかしい。	お化け話，「開かずの扉」，恐怖物語，久原房之介
9	太宰治の小説である「女生徒」で午後の図画の時間に「私」をモデルにしたのは誰ですか？	午後の図画の時間には、皆、校庭に出て、写生のお稽古。伊藤先生は、どうして私を、いつも無意味に困らせるのだろう。きょうも私に、先生ご自身の絵のモデルになるよう言いつけた。三十分間だけ、モデルになってあげることを承諾する。すこしでも、人のお役に立つことは、うれしいものだ。けれども、伊藤先生と二人で向かい合っていると、とても疲れる。それからずいぶん自分を買いかぶっているのですよ、ああ、こんな心の汚い私をモデルにしたりなんかして、先生の画は、きっと落選だ。美しいはずがないもの。いけないことだけれど、伊藤先生がばかに見えてしようがない。	伊藤先生
10	太宰治の小説である「女生徒」の主人公が机の上に飾っている花の名前は？	お部屋へ戻って、机のまえに坐って頬杖つきながら、机の上の百合の花を眺める。	百合
11	太宰治の小説である「女生徒」で主人公が朝、掃除をしながら歌った曲は？	お掃除しながら、ふと「唐人お吉」を唄う。普段、モオツァルトだの、バッハだのに熱中しているはずの自分が、無意識に、「唐人お吉」を唄ったのが、面白い。蒲団を持ち上げるとき、よいしょ、と言ったり、お掃除しながら、唐人お吉を唄うようでは、自分も、もう、だめかと思う。	唐人お吉
12	太宰治の小説である「女生徒」で、「私」が新聞で一番楽しいと感じている内容は？	新聞では、本の広告文が一ばんたのしい。	新聞の本の広告文
13	太宰治の小説である「女生徒」で、主人公がおととしの夏休みに遊びに行った場所は？	おととしの夏休みに、北海道のお姉さんの家へ遊びに行ったときのことを思い出す。苫小牧のお姉さんの家は、海岸に近いゆえか、始終お魚の臭いがしていた。	北海道，苫小牧
14	太宰治の小説である「女生徒」で「私」が見たいと思っている映画は？	お風呂から上がって、私と二人でお茶を飲みながら、へんにニコニコ笑って、お母さん何を言い出すかと思ったら、「あなたは、こないだから『裸足の少女』を見たい見たいと言ってたでしょう？　そんなに行きたいなら、行ってもよござんす。そのかわり、今晩は、ちょっとお母さんの肩をもんで下さい。働いて行くのなら、なおさら楽しいでしょう？」もう私は嬉しくてたまらない。「裸足の少女」という映画も見たいとは思っていたのだが、このごろ私は遊んでばかりいたので、遠慮していたのだ。	「裸足の少女」
15	太宰治の小説である「女生徒」で、「私」の隣の席の人の名前は？	キン子さんは、全く無性格みたいで、それゆえ、女らしさで一ぱいだ。学校で私と席がお隣同士だというだけで、そんなに私は親しくしてあげているわけでもないのに、お寺さんのほうでは、私のことを、あたしの一ばんの親友です、なんて皆に言っている。	キン子