クエリを考慮した新規手法により関連する企業データを減らしLLM APIの使用コストを最適化

AREFEEN Md Adnan　DEBNATH Biplob　CHAKRADHAR Srimat

要　旨

ユーザークエリに対し、より正確な回答を大規模言語モデル（LLM）に生成させるために企業所有のデータをコンテキストとして使用すると、LLM APIの使用コストは急激に上昇します。本稿では、こうしたコストを削減するために開発したLeanContextを紹介します。このシステムは、関連する企業データのコンテキストから、クエリを考慮し、コンパクトでAIモデルと親和性の高い要約を生成します。これはクエリを考慮せずに、人間に優しい要約を生成する従来のツールとは異なります。最初に、検索拡張生成（RAG）を使用して、重要でクエリに関連した企業データを含む、クエリを考慮した企業データのコンテキストを生成します。次に、強化学習を用いてコンテキストを更に縮小します。同時に、ユーザークエリと縮小コンテキストからなるプロンプトにより、元の企業データコンテキストを使用したプロンプトと同程度に正確な応答をLLMから引き出せるようにします。この縮小コンテキストは、クエリに依存するだけでなく、サイズの変更が可能です。実験の結果から、LeanContextは、(a) LLM応答の精度を維持しながらLLM API使用コストを（RAGとの比較で）37%～68%削減でき、(b) この最先端の要約ツールでRAGコンテキストを縮小すると応答の精度を26%～38%向上できることが明らかになりました。

KeyWords　生成AI／大規模言語モデル／テキスト要約／自然言語処理／強化学習／ドメイン固有質問応答／検索拡張生成

1. はじめに

大規模言語モデル（Large Language Model、以下、LLM）は、人間のような言語を生成するために広範囲のテキストデータに基づいて訓練された先進的なAIモデルであり、自然言語処理によるタスクを大幅に強化します。著明な例であるOpenAIのGPT-4[1]は、ユーザーに優しいアプリケーションプログラミングインタフェース（API）を特長としており、これがコンテキスト対応チャットボットやリアルタイム言語翻訳、効率的なテキスト要約といった分野での幅広い使用を促しています。これにより、さまざまな業界においてユーザーエクスペリエンスの向上が図られています。

GPT-4のようなLLMは、企業が所有するデータに関する情報についてのクエリに答えることができません。というのも、LLMがこうしたデータで訓練されていないからです。しかし、LLMに企業の所有データを認識させると、業界に特有の用語やプロセス、コンテキストを使用した回答を生成することができます。これにより、企業はより正確で関連性の高い回答を得られるようになります。

LLMに企業データを認識させるためのよく知られた2つの方法は、ファインチューニングと検索拡張生成（RAG）[2]です。ファインチューニングは、LLMのモデルの重みを変更して、モデルをドメイン固有の微妙な違いに適応させます。それとは対照的に、RAGは事前学習済みのモデル（修正はまったく加えない）と、企業データから関連情報（コンテキスト）を選び出すリトリーバーの組み合わせを利用して、この外部知識をLLMへのプロンプトに組み込みます。本研究では、LLMに企業データを認識させるためにRAGアプローチを使用しています。

LLM APIの使用コストは、特にLLMへのプロンプトに企業情報を組み込む場合、非常に迅速に増加する可能性があります。コストは、プロンプトとLLMの応答に含まれるトークン数に依存します。GPT-3 LLMではトークンは約4文字[3]ですが、LLMや言語によってこの数字は異なります。

LLM API使用の高コストを示す具体例として、週に2回、15,000人の訪問者がそれぞれ3つのリクエストを送るサービスを考えてみます。（典型的な）プロンプト1つ当たりでは、約1,800のプロンプトトークンと80の出力トークンがあるとします[4]。この場合、GPT-4 APIの使用コストは月額21,200ドルになります（価格は、プロンプトに対して0.03ドル/1Kトークン、生成された出力に対して0.06ドル/1Kトークンとします）。

本稿では、企業データの使用がより有用な回答を生成できるシナリオにおいて、LLM APIの使用コストの削減に注目します。そして、優れたコスト効率でクエリアウェアに企業データのコンテキスト検索を可能にする新しいシステム「LeanContext」を提案します。これにより検索されるコンテキストはコンパクトになり、回答のためのクエリとの関連性が高まります。実験の結果は、(a) LeanContextなら、応答の高い精度を維持しながらLLM API使用コストを削減（RAGコンテキストと比較して37%～68%）できること、及び (b) 前述したよく知られる要約ツールでRAGコンテキストを縮小するケースと比べて、LeanContextは応答の精度を26%～38%向上させることを示しています。

2. 検索拡張生成

図1は、従来のRAG方式を示しています。これは、企業データの取り込み、及びクエリ/応答という、並行して動作可能な2つの異なるパートで構成されます。

企業データの取り込み：企業のテキスト文書はテキストスプリッタで小さなチャンクに分割されます（チャンクとは、企業文書内の連続した文の集まり）。埋め込みジェネレータは、各チャンクをあらかじめ決められたn次元のベクトルに埋め込みます。これらのベクトル及び対応するチャンクは、ベクトルデータベースに保存されます。チャンクをベクトルとして保存することで、特定のユーザークエリに関連する企業データを容易に見つけることが可能となります。

クエリ/応答：ユーザークエリもn次元のベクトルに埋め込まれます。図1に示したように、セマンティック検索の手法は、ベクトルデータベース内でユーザークエリベクトルに最も似ているN個のベクトルを特定します。ここで、Nは事前に決定したパラメータです。これらN個のベクトルに対応するチャンクは関連する「RAGコンテキスト」となります。このコンテキストはユーザーのクエリと組み合わされてプロンプトが構築され、LLMからの応答がユーザーに提供されます。プロンプトのサイズはLLM APIの最大トークンの制限を超えることができません（この制限はLLMによって変わる可能性があります）。

3. LeanContext

LeanContextを設計するうえで鍵となる2つの重要な見解があります。1つ目は、実際のアプリケーションでの実験から、LLMが正確な応答を生成する際に、従来のRAGで検索されたコンテキストのN個のチャンクにある情報すべてを必要とするわけではないということが分かっています。2つ目は、RAGコンテキスト内のどの特定の情報が省略可能かは、「ユーザークエリ」に依存しているということです。LeanContextは、これらの所見を活用して、RAGコンテキストからよりコンパクトな「縮小コンテキスト」を構築します。この縮小コンテキストはLLM API使用コストの削減とダイレクトにつながっています。また、縮小コンテキストは、より大きな（N個のチャンクの）RAGコンテキストによるプロンプトを使った場合と同程度の精度を持つLLM応答を実現します。

図2にLeanContextのシステム概要を示します。まず、従来のRAG手法を使用して、企業データコンテキス

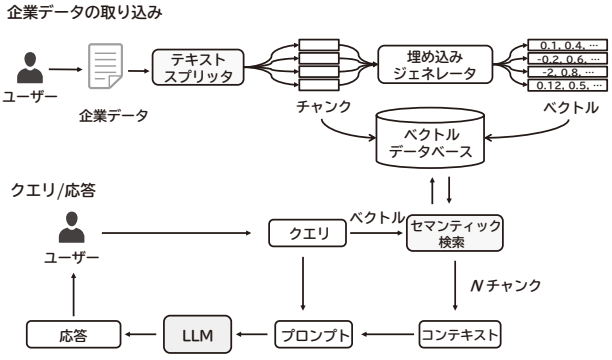

図1　検索拡張生成

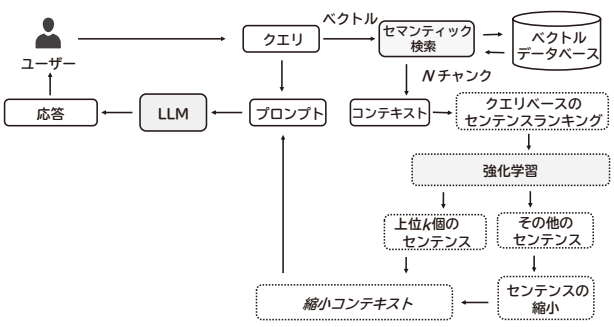

図2　LeanContextのシステム概要

トのN個のチャンクを検索します。次に、このRAGコンテキスト内のセンテンスをユーザークエリとの関連性に基づきランク付けをします。第3章1節の新しい強化学習アルゴリズムでは、縮小コンテキストの構築のために考慮する必要があるランク付けされたRAGコンテキストのなかから、上位k個のセンテンスを特定します。本研究の強化学習アルゴリズムは、ユーザークエリとRAGコンテキストに基づいてkの値を決定します。その後、重要な上位k個のセンテンスはそのまま残されますが、ランク付けされたRAGコンテキスト内にあって重要度が低いその他の文は、フレーズ削除あるいは要約によって更に圧縮されます。そして、第3章2節で説明するように、上位k個のセンテンスと圧縮された重要度の低いセンテンスを組み合わせて縮小コンテキストを作成します。

3.1 kを算出するための強化学習

ユーザークエリとこれに対応するRAGコンテキストを与えると、NECの「新規軽量Q学習ベースの強化学習(RL)」アルゴリズムは、この組み合わせに適切なkを計算します。次に、このRLアルゴリズムの状態、行動、報酬のコンポーネントについて簡単に説明します。

状態：N個のチャンクからなるRAGコンテキストの埋め込みベクトルを作成します。次に、RAGコンテキスト埋め込みベクトル（v_c）からクエリ埋め込みベクトル（v_q）を引いて差分ベクトルを得ます。多数のクエリとコンテキストのペアに対して差分ベクトルを構成し、これらのベクトルをクラスタリングしてセントロイドを算出します（K-Meansクラスタリングアルゴリズムを使用）。これらのセントロイドが本研究で用いる状態ベクトル$\mathcal{S}$です。

$$\mathcal{S} \leftarrow \textit{K-Means}\left(\bigcup_{i,j}\left(v_{c_i} - v_{q_j}\right)\right)$$

変数$_i$と$_j$は、それぞれ異なるRAGコンテキストとユーザークエリにインデックス付けするため使用します。

行動：行動はランク付けされたRAGコンテキストに含まれる総センテンス数の特定の割合に対応しています。行動は0～0.4の間で任意の値をとることができ、0.05刻みになっています。例えば、行動が最大値の0.4であったとするなら、ランク付けされたRAGコンテキストに含まれる総センテンス数の40%がクエリに最も似ている上位k個のセンテンスであるとみなせるでしょう。上位k個のkは、現在の行動値及びランク付けされたRAGコンテキストに含まれる総テキスト数の積として導き出せます。

報酬：行動の値が与えられると、上位k個のセンテンスとそのトークン数を割り出すことができます。トークン比（τ）は、縮小コンテキストに含まれる上位k個のセンテンスのトークン数と、ランク付けされたRAGコンテキストに含まれるすべてのセンテンスのトークン数の比として定義されます。トークン比が低いほど、上位kコンテキストの長さは短くなります。しかし、上位k個の縮小コンテキストに対するLLMの応答精度は、完全なRAGコンテキストに対するLLMの応答精度と同等でなければなりません。異なるLLMでの応答精度の比較には、ROUGE-1[5]スコアを使用します（ROUGE-1スコアを計算するために使用する参照応答は、第3章3節で説明します）。完全なRAGコンテキストでのROUGEスコアが（r^*）で、上位k個コンテキストのROUGEスコアがrである場合、（τ） $r - r^* \geq 0$であるならQテーブル内の現在の（状態、行動）ペアの値には報酬を与えられます。そうでない場合にはペナルティーが課されます。報酬関数Rは次のように定義します。

$$\boldsymbol{R = -(1-\alpha)\tau + \alpha(2r - r^*)}$$

ここで、αはトークン比と応答精度の、報酬値に関する相対的な寄与を制御します。

3.2 縮小コンテキスト

開発したRLアルゴリズムは、クエリに固有のkの値を割り出します。これにより上位k個のコンテキストが決まります。このコンテキストにはクエリに関連する重要な上位k個のセンテンスに加えて、上位k個のセンテンスの近傍にあるその他の重要度の低いセンテンスが含まれます。**図3**

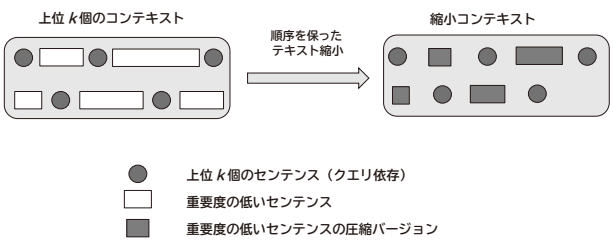

図3　今回の縮小コンテキストの構成

に今回の縮小コンテキストの構成方法を示します。クエリに対するコンテキストの関連性を維持するのに重要であるため、最も関連性の高い上位k個のセンテンスはそのまま残します。しかし、重要度の低いセンテンスは、オープンソースのテキスト縮小手法[6) -10)]を使用して更に個別に圧縮します。また、最後の上位k個を超えるセンテンスは縮小コンテキストに含めません。

ランク付けされたRAGコンテキストにおいて、上位k個のセンテンスと重要でないセンテンスの両方で元の順序は保持されます。センテンスの順序を保持することによりコンテキストの時間的な一貫性が確保されます。縮小コンテキストの構成にこうした全体的なアプローチをとることにより、LLM APIの使用コストを大幅に削減しながら、LLM応答の精度の確保を最終的に実現します。

3.3 結果

企業データ：本研究ではarXivとBBCニュースのデータリポジトリを使用しました。これらのデータリポジトリには2023年3月に公開された文書[6)]が含まれており、GPT-3.5-Turboモデルの事前訓練には使用されていません。まず、arXivからはランダムに25の文書を選びました。これらの文書には63～962のセンテンス、1文書当たり平均で352のセンテンスがあります。同様に、BBCニュースからは100の文書を選びました。これらの文書には4～139のセンテンス、1文書当たり平均で30のセンテンスがあります。

クエリと参照応答：本研究ではQAGenerationChain[11)]を用いて、各データセットでそれぞれ100個のクエリを生成しました。これらは、RAGやLeanContextのROUGE-1精度スコアを算出するための「基準」となるLLM応答を得るために、文書全体をコンテキストとして使用します。

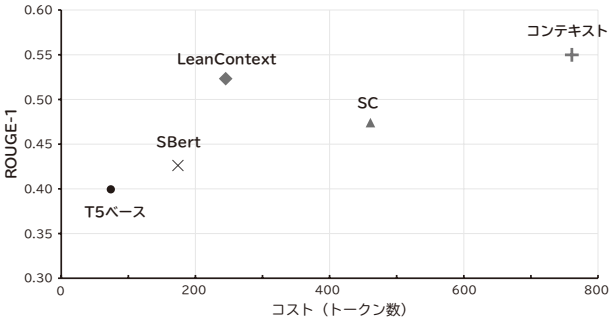

図4　LeanContextと有名な要約ツールでRAGコンテキストの縮小を行ったときの比較

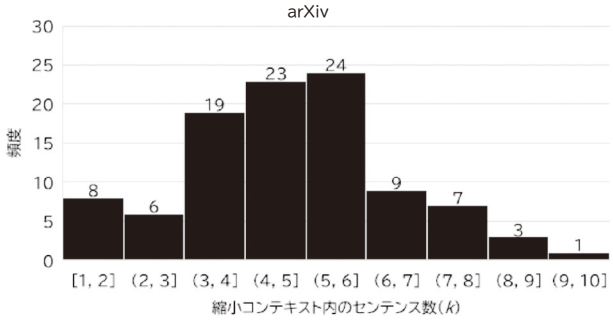

図5　arXivデータでの（重要なセンテンス）kの分布

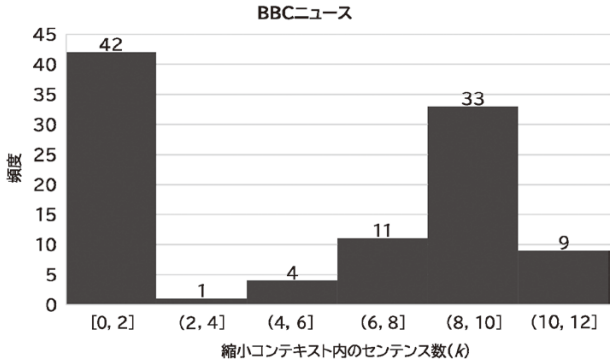

図6　BBCニュースデータでの（重要なセンテンス）kの分布

企業データコンテキスト：arXivデータセットに対してはN=4、BBCニュースに対してはN=8を設定しました。arXivデータセットの場合、RAGコンテキストの総センテンス数は9～25の間で変動し、コンテキスト当たりの平均センテンス数は15でした。同様に、BBCニュースデータの場合、コンテキスト内の総センテンス数は18～34の間で変動し、コンテキスト当たりの平均センテンス数は26でした（**図4**）。

縮小コンテキスト：arXivデータとBBCニュースデータにおける上位k個のセンテンスの分布を、それぞれ**図5**と**図6**に示します。

表1では、RAGコンテキストと縮小コンテキストとの効果を比較しています。LLMの応答精度（ROUGE-1スコア）はほぼ同じですが、縮小コンテキストはLLM APIの

表1　従来のRAGと縮小コンテキストとの比較（精度はROUGE-1スコア）

データ	コンテキスト	精度	トークン数（プロンプト、応答）	コスト削減
arXiv	RAG (*N*=4)	39.85	521, 26	-
	縮小	38.44	321, 22	37%
BBC	RAG (*N*=8)	54.98	724, 37	-
	縮小	52.33	218, 27	68%

表2　他の有名な要約ツールでRAGコンテキストを縮小する際の応答精度の向上

データ	要約ツール	ROUGE-1	改善
BBCニュース	SBert	42.61	-
	SBert+ LeanContext	53.55	26%
	T5	39.93	-
	T5 + LeanContext	55.32	38%

使用コストを37%～68%削減します。T5[10]、BERT[9]、SC[6]などの有名なテキスト縮小モデルと比較すると、LeanContextはコストを削減し、精度が向上します（図4）。また、他の有名な要約ツールを使用してRAGコンテキストを縮小するケースでも、応答精度（ROUGE-1スコア）が向上します（**表2**）。

4. むすび

LeanContextは、高い精度を維持しながら、LLM APIの使用に関係したコストを軽減する、コスト効率に優れたクエリ考慮型コンテキスト縮小システムです。コンテキストの縮小はLLMの推論時間の改善にもつながります。LeanContextは、他の有名な要約ツールと組み合わせてRAGコンテキストを縮小するケースでも、効果的に使用できます。

参考文献

1) OpenAI: GPT-4
https://openai.com/product
2) Patrick Lewis et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 34th Conference on Neural Information Processing Systems (NeurIPS), 2020
https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
3) Claudia Slowik: How Much Does It Cost to Use GPT Models? GPT-3 Pricing Explained, Neoteric, 2023.2
https://neoteric.eu/blog/how-much-does-it-cost-to-use-gpt-models-gpt-3-pricing-explained/
4) Lingjiao Chen, Matei Zaharia and James Zou: FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance, 2023
https://arxiv.org/abs/2305.05176
5) Chin-Yew Lin: ROUGE: A Package for Automatic Evaluation of summaries, The Workshop on Text Summarization Branches Out (WAS), pp.74-81, 2004
https://aclanthology.org/W04-1013/
6) Yucheng Li: Unlocking Context Constraints of LLMs: Enhancing Context Efficiency of LLMs with Self-Information-Based Content Filtering, 2023
https://arxiv.org/abs/2304.12102
7) Md Tahmid Rahman Laskar, Mizanur Rahman, Israt Jahan, Enamul Hoque and Jimmy Huang: CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia, 2023
https://arxiv.org/abs/2305.06147
8) Henry Gilbert, Michael Sandborn, Douglas C. Schmidt, Jesse Spencer-Smith and Jules White: Semantic Compression With Large Language Models, 2023
https://arxiv.org/abs/2304.12512
9) Nils Reimers and Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp.3982-3992, 2019
https://aclanthology.org/D19-1410/
10) Colin Raffel et al.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research volume21, 2020
https://jmlr.org/papers/volume21/20-074/20-074.pdf
11) GitHub: LangChain
https://github.com/hwchase

執筆者プロフィール

AREFEEN Md Adnan
NEC Laboratories America
Research Assistant
University of Missouri-Kansas City

DEBNATH Biplob
NEC Laboratories America
Senior Researcher

CHAKRADHAR Srimat
NEC Laboratories America
Department Head

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報（日本語）　NEC Technical Journal（英語）

Vol.75 No.2　ビジネスの常識を変える生成AI特集
～社会実装に向けた取り組みと、それを支える生成AI技術～

ビジネスの常識を変える生成AI特集によせて
生成AI技術への取り組み ～基盤から応用、ルール作りまで～

◇ 特集論文

急速に広まる生成AIの市場適用
NECにおける生成AIの取り組みについて
電子カルテと医療文書の作成支援による医師業務効率化
映像×LLMによる報告書作成業務の自動化
映像分析と生成AIによるリアル世界の行動理解
サイバー脅威インテリジェンス生成自動化
生成AIの社内活用を進めるNEC Generative AI Service（NGS）
ソフトウェア・システム開発への生成AIの活用
LLMとMIで革新する素材開発プラットフォーム
LLMと画像分析を活用した被災状況の把握

生成AIの可能性を高める基盤技術
日本語性能に優れたNECの大規模言語モデル（LLM）
NECの生成AIを支える国内企業で最大規模のAIスーパーコンピュータ
より安全な大規模言語モデル（LLM）を目指して
データを秘匿したまま連携できる連合学習技術とLLMへの適用可能性
大規模言語モデル（LLM）によるFew-shotクラスタリングの可能性
オープンドメイン常識推論のための知識拡張型プロンプト学習
AI連携とオーケストレーションのための基盤ビジョンLLM
クエリを考慮した新規手法により関連する企業データを減らしLLM APIの使用コストを最適化

AI技術が社会へ浸透するために
AI標準化・ルールメイクに関する動向とNECの取り組み
人権尊重に向けたNECのAIガバナンスの取り組み
RCModelを用いたAIリスクマネジメントのための人材育成事例

◇ NEC Information
2023年度C&C賞表彰式典開催

Vol.75 No.2
（2024年3月）

特集TOP