

地方議会の予算表を対象としたLLMによる表形式変換を用いたRAGの提案

前多陸玖(小樽商科大学) 木村泰知 (小樽商科大学)

背景・目的

RAGにおける課題

- 文書にHTMLや表・画像等が含まれていると出力の精度が低下する
- embedding modelやchunk sizeなどに左右される

本研究の目的

- M-LLMを用いた表構造理解の手法を提案する
- 従来手法と提案手法においてどのような差があるのかを検証した
- 複数のembedding modelやchunk size, overlap sizeにおいてそれぞれの差を比較した

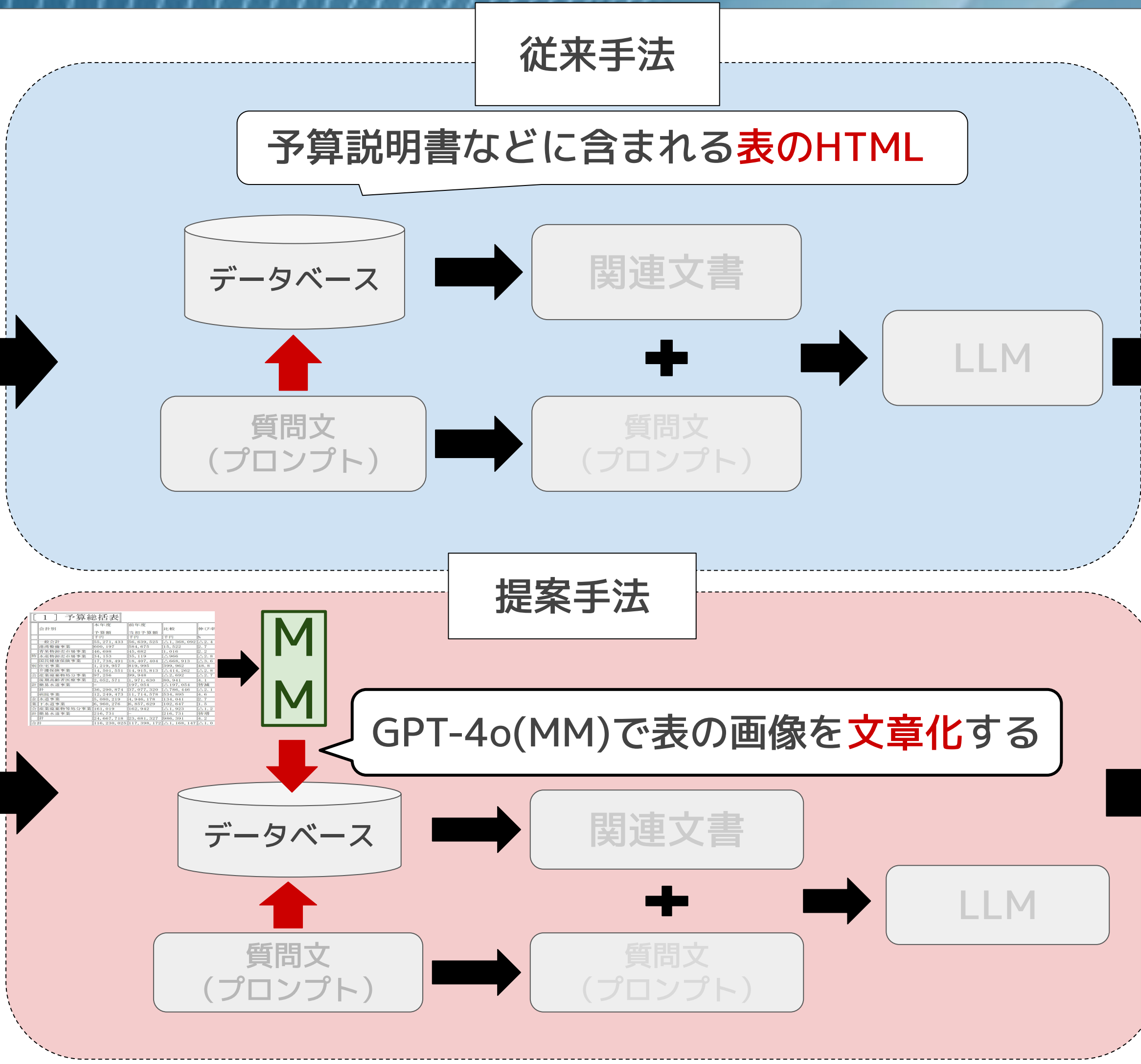
実験方法

発言文	
otaru_r04-01-sent28	本市はこれまで、将来にわたって効率的かつ安定的に持続可能な行政運営を…
otaru_r04-01-sent29	とりわけ、市税などの歳入動向がコロナ禍により予測しにくい現状においては…
otaru_r04-01-sent32	その結果、行政経費は自治体DXの推進に係る経費などで増加しましたが…
⋮	
otaru_r04-01-sent64	小樽港につきましては、第3号ふ頭において、クルーズ船受入環境の充実を…
otaru_r04-01-sent65	また、港湾機能の保全を図るため、老朽化した北防波堤や色内ふ頭護岸の…

実験対象

MBLinkに含まれる小樽市の令和4年度の議会会議録と予算表から

- 発言文…46文
- 予算表…46table



出力		
otaru_r04-tab166	### 予算科目と予算額 #### 2小樽ファンが支えるふるさと…	0
otaru_r04-tab131	入力された情報に基づいて、財政調整基金に関連する情報を検索しました…	0
otaru_r04-tab256	### 財政調整基金費 - **本年度予算額**: 749千円	1
⋮		
otaru_r04-tab108	### 第3号ふ頭及び周辺再開発事業 - **第3号ふ頭岸壁改良事業**: 134,000	1
otaru_r04-tab146	### 港湾事業債 - **北防波堤改良事業**: 15,000千円	0
出力		
otaru_r04-tab21	1. **市税** - 本年度予算額: 13,815,300千円	1
otaru_r04-tab131	入力された内容に関連する情報として、以下のデータが該当します…	0
otaru_r04-tab256	- **本年度予算額**: 288,398千円 - **前年度予算額**: 298,453千円	1
⋮		
otaru_r04-tab108	1. **第3号ふ頭岸壁改良事業** - **金額**: 18,300千円	1
otaru_r04-tab612	### 北防波堤改良事業 - **本年度予算額**: 15,000千円	1

実験①「従来手法との比較」

実験①の目的

- 提案手法と従来手法における性能の違いを検証する

embedding model	text-embedding-ada-002
chunk size	200
overlap size	0
評価	正解率を用いる

実験②「パラメータの比較」

実験②の目的

- embedding modelやchunk size、overlap sizeにおけるそれぞれの性能の違いを検証する

embedding model	text-embedding-ada-002、3-large、3-small
chunk size	200、400、600、800、1000
overlap size	chunk sizeに対してそれぞれ0%、25%、50%
評価	正解率を用いる

従来手法との比較結果

chunk size	提案手法	従来手法
200	54.35%	41.30%
400	52.17%	41.30%
600	56.52%	41.30%
800	56.52%	41.30%
1000	54.35%	41.30%

- 提案手法ではどのchunk sizeにおいても55%近い正解率を算出したのに対して、従来手法ではどのchunk sizeにおいても正解率は41.30%であった
- chunk sizeが600、800の場合において正解率が56.52%と最大で15%以上の性能改善が見られた

パラメータの比較結果

model	text-embedding-ada-002			text-embedding-3-large			text-embedding-3-small		
chunk size	overlap size0%	overlap size25%	overlap size50%	overlap size0%	overlap size25%	overlap size50%	overlap size0%	overlap size25%	overlap size50%
200	54.35	56.52	47.83	45.63	45.65	41.30	58.70	63.04	56.52
400	52.17	50.00	50.00	47.83	43.48	52.17	52.17	56.52	58.70
600	56.52	43.38	47.83	41.30	36.29	32.61	56.52	50.00	50.00
800	56.52	50.00	47.83	45.65	43.48	34.78	54.35	50.00	50.00
1000	54.35	50.00	50.00	45.65	47.83	39.13	52.17	54.35	56.52

- 正解率が最も高かったのは、embedding modelがtext-embedding-3-smallでchunk sizeが200、overlap sizeが25%の時で63.04%であった
- それぞれの分析において、embedding modelではtext-embedding-3-smallが最適でoverlap Sizeは検索に大きな影響を与えていないことがわかった

考察

分類	合計
全てのケースで正解	12
全てのケースで不正解	14
提案手法のみ全て正解	5
従来手法のみ全て正解	1
提案手法のみ一部正解	8
従来手法のみ一部正解	0
どちらのケースでも正解	6

提案手法の優位性

提案手法のみで正解できた件数は合計で13件、従来手法のみで正解できた件数は1件であった。

提案手法により、従来手法では正解することのできないケースにアプローチすることができた

提案手法におけるチャンク例

2. 後期高齢者医療広域連合納付金
- **不年度予算額**: 2,179,281千円
- **前年度予算額**: 2,224,870千円
- **比較**: △45,589千円
- **各目節**: 10
- **明細細金額**: 2,179,271千円

チャンク1

従来手法におけるチャンク例

予算科目 1保険料
本年度予算額 2,658,742
前年度予算額 2,651,246
比較 7,496
節 1介護保険料
説明 2,658,742
区分 2,651,246
金額 7,496

チャンク1

チャンク2

チャンクの違い

提案手法のチャンクではひとつのチャンクにおいて、表のHeaderとそれに対応する値が隣り合っていることがわかる
一方で、従来手法のチャンクでは表のHeaderとそれに対応する値が、チャンク1とチャンク2で分離している

まとめ

本研究ではM-LLMを用いた表形式変換によるRAGの改善を提案した。従来手法に対して提案手法は、正解率が15%以上改善した