被引用統計データのセル特定データセットの構築

中野優[1]　加藤誠[2]

本論文では文書中の数値の真偽を検証するために，数値が参照する統計データのセルを自動的に特定する問題を提案する．我々はこの統計データのセル特定問題をアドホック検索問題と見なし，数値と数値を含む文書の組をクエリ，統計データのセルを検索対象の文書として問題を定式化した．またセル特定問題のためのデータセットとして，Wikipedia と統計データへのリンクを用いて日本語と英語の 2 つの言語からなるデータセット WikiStatCells を作成した．さらに本問題の難しさを検証するために，ニューラル言語モデルを含む複数の検索手法を実装し，作成したデータセットを用いて実験を行った．

1　はじめに

文章を書く際において，数値的な情報の根拠として統計データを引用することは多い．このような引用は**統計データ引用**と呼ばれ，論文においても統計データ引用は年々増加傾向にある [15]．しかしながら，統計データは正しく引用されない場合 (誤引用) が存在する．我々のこれまでの研究によると，Wikipedia における統計データ引用において，記述されている数値と被引用統計データ内の数値が異なる例が複数存在することが判明している [20]．

このような状況に対して，我々はこれまでの研究において，文章からどの統計データが引用されているかを特定する研究を行ってきた [20]．この研究においては，文章から統計データを検索し，文章に関連する統計データのランキングを提示している．しかしながら誤引用をより詳細に検証するためには，提示された統計データの中から文章中の数値に対応するセルを特定する作業が必要となる．我々の調査によると大半の統計データは数百から数万という多くのセルを持つため，セルを特定するためには多くの時間がかかるという問題がある．

そこで本論文では文章中の数値から統計データ中のセルを特定する問題 (**セル特定問題**) を提案する．この問題においては，数値とその数値を含む文書に加えて，文書において引用されている統計データが与えられたときに，数値に対応する統計データ中のセルを特定する．本論文ではこの問題をアドホック検索の問題として定式化する．つまり，数値と数値を含む文書の組をクエリとみなし，統計データに含まれる全てのセルを検索対象とみなして検索を行う問題として定式化する．

さらに本論文では，セル特定問題のためのベンチマーク用データセット WikiStatCells を作成した．このデータセットは Wikipedia に含まれる数値に対して，その数値に対応する統計データのセルがアノテーションされたデータセットである．本論文ではこのデータセットを，Wikipedia の記事と記事からリンクされている統計データを用いてクラウドソーシングでアノテーションを行い，日本語と英語の 2 つの言語のデータセットを構築した．

本問題の難しさを検証するために，作成したデータセットを用いて実験を行った．特に，クエリとしてどの情報を用いるべきかについての比較と，ニューラル言語モデルを用いた表質問応答手法を含む複数の手法についての性能の比較を行った．その結果，数値周辺の単語だけでなくタイトルもクエリとして用いた場合の方が性能が向上することが判明した．また，表質問応答手法は fine-tuning をしない状態であっても，一部の基本的な検索手法を上回る性能を発揮することが判明した．

本論文の貢献は以下のとおりである．

1. 文章中の数値から被引用統計データのセルを特定する問題を提案し，この問題をアドホック検索問題として定式化した．
2. 統計データに対するセル特定問題のための新たなベンチマーク用データセットを構築した．
3. 構築したデータセットを用いた実験により，複数の検索手法の比較検証を行った．

本論文の構成は次の通りである．第 2 節では関連研究として文章中の数値からセルを対応付ける研究と，表を対象とした質問応答の研究について説明する．第 3 節ではセル特定問題の問題設定と定式化について説明する．第 4 節では構築したセル特定問題のためのデータセットである WikiStatCells について説明する．第 5 節では構築したデータセットを用いた評価実験について説明し，第 6 節では本論文の結論とともに今後の課題について説明する．

2　関連研究

本節ではセル特定問題関連する研究として，2.1 節で類似する問題設定を持つ，文とセルを対応付ける研究を紹介した後，2.2 節で表検索の研究について紹介し，2.3 節で関連する問題設定である表質問応答の研究を紹介する．

2.1　文とセルの対応付け

文章において，表を用いて説明を行うことは一般的である．これに対して，文章とその文章に含まれる表のセルを対応付ける研究が行われている [7, 9]．Kim らは論文や金融関連の報告書などの PDF の文書において，表を参照するような文を特定し，その文が参照するセルをハイライトして文とともに提示するシステムを提案している [9]．また Kim らはこのシステムを用いたユーザスタディを実施し，文とセルをハイライトした表を提示することで，ユーザは文章とセルの対応関係をより正確かつ素早く理解できるようになることを示している．また，Ibrahim らは，Web 上の文書と付随する表に対して，文書中の数値と表のセルを対応付ける手法を提案している [7]．

これらの研究は文章と表のセルを対応付けるという点において，本論文のセル特定の問題設定と類似している．しかしながら

[1] 学生会員　筑波大学大学院人間総合科学学術院
s2030514@s.tsukuba.ac.jp
[2] 正会員　筑波大学図書館情報メディア系 / JST さきがけ
mpkato@acm.org

これらの研究は，文章中に含まれる表に対してセルを対応付ける研究である．よって，表のサイズが文章に収まる程度に小さいことが多く，また表周辺の文のみからセルを対応付けることが可能である．一方で，本論文は引用されている外部の表に対してセルを対応付ける研究である．そのため，表のサイズはこれらの研究と比較して大きいことが多く，文書のタイトルなどの文以外の情報も対応付けに必要となるという点において，上記の研究とは異なる．

2.2 表検索・統計データ検索

統計データに類似するアイテムとして，表を対象とした検索の研究を説明する．アドホック表検索タスクはテキストをクエリとして，クエリに適合する Web 上の表を検索するタスクである [1, 3, 11–13, 18]． Zhang と Balog はアドホック表検索タスクのデータセットである WikiTables データセットを提案し，さらにランキング学習を用いて表を検索する手法や単語埋め込みとエンティティ埋め込みを用いた Semantic Matching によって表を検索する手法を提案している [18]．アドホック表検索タスクにおいては Zhang と Balog が提案した WikiTables データセットを対象に研究が行われており，パッセージ検索と多様体学習を用いる手法 [11]，行列分解を用いる手法 [1]，マルチモーダル深層学習を用いる手法 [12]，BERT を用いる手法 [3]，グラフニューラルネットワークを用いた手法 [13]など様々な手法がこれまでに研究されてきた．既存の表検索データセットで対象となる Web 上の表は，4.2 節で説明する通り，表のサイズが小さい点や数値セルの割合が大きい点など，本論文が扱う統計データとは異なる点が存在する．

また，テキストから統計データを対象として検索を行う研究も存在する [2, 8, 20, 21]． Kato らは日本とアメリカの政府統計データポータルの統計データを収集し，アドホック統計データ検索のベンチマーク用データセットを提案した [8]． Chen らは，統計データから抽出した列名を検索に用いることで，統計データの検索性能を向上させる手法を提案した [2]．さらに，data.gov の統計データを対象として 6 つの検索タスクを設定して統計データ検索用のデータセットを構築し，ベースライン手法と比較して提案手法が良い検索結果を提示できることを示した．岡本と宮森は統計データにカテゴリを付与し，検索時にクエリからカテゴリを絞り込むことで統計データの検索性能を向上できることを示した [21]．中野と加藤は Wikipedia の文章をクエリとして統計データを検索する際に，クエリとなる文章と統計データの両方のフィールドを利用することで検索性能を向上できることを示した [20]．これらの研究は与えられたテキストと多数の表の集合の中からテキストに対応する表を検索するタスクである一方で，本論文はテキストと 1 つの表からテキストに対応する表のセルを特定するタスクであり，問題設定が異なる．

2.3 表質問応答

表質問応答は与えられた表と質問に対して，表の情報を用いて与えられた質問に回答するタスクであり，数多くの研究が取り組まれてきた [5, 6, 10, 17, 19]．表質問応答における質問は，表の 1 つのセルが答えとなる Lookup 質問と，表の複数のセルを集約した結果が答えとなる Aggregation 質問の 2 つに分けられる [5]．これまでの多くの表質問応答の研究においては，後者の Aggregation 質問に対応するために Semantic Parsing を用いた手法が提案されてきた [10, 17]． Semantic Parsing とは与えられた文を論理形式に変換するタスクであり，表質問応答においては与えられた質問を SQL などの論理形式に変換し，その論理形式を表上で実行することで質問の回答を行う． Yin らは，Web 上の表と周辺テキストから事前学習を行う TaBERT を提案し，TaBERT によって得られた質問や表の表現ベクトルを既存の深層学習を用いた Semantic Parsing 手法の入力とすることで，表質問応答の性能を改善した [17].

しかしながら深層学習を用いた Semantic Parsing 手法は，正解となる論理形式が教師データとして必要であり，大規模なアノテーションにはコストがかかるという問題がある．これに対して，BERT [4] などのニューラル言語モデルと弱教師あり学習を用いて質問の回答に用いられるセルの確率と用いられる演算を予測することによって，論理形式を経由せずに質問の回答を作成する手法が提案されている [5, 6]． Herzig らは，BERT [4] をベースとして Wikipedia の表と周辺のテキストを元にした事前学習を行い，質問と答えのみから回答に用いられるセルと演算を予測する Fine-tuning を行うことによって，論理形式を経由しない表質問応答手法を提案した [6]． Glass らは，質問から回答に使われる行と列をそれぞれ予測し，そこから回答に使われるセルを求める手法を提案している [5]．この手法は特に Lookup 質問に対して高い性能を示し，WikiSQL データセット [19] のテストデータにおける Lookup 質問において 97.99% という高い正解率を達成した．

表質問応答のうち，特に Lookup 質問は，文 (質問) から対応するセルを探すという点において本研究と類似している．特に，Semantic Parsing を用いない表質問応答手法は，文 (質問) に対応するセルを確率として予測するため，本論文の問題設定に適用可能である．具体的には，検証対象の数値が含まれる文と表を入力として，表の各セルの確率を計算し，確率が高い順にセルを並べ替えることで，セルのランキングが作成できる．しかしながら，本論文の入力は質問ではなく文であることと，本論文のセル特定においては数値周辺の文だけでなくタイトルなど別の情報も必要となること，本論文が扱う統計データは表質問応答が扱う Web 上の表と比較してサイズが大きいという違いが存在する．

3 セル特定問題

本節ではセル特定の問題設定について説明する．セル特定タスクとは，文書中の数値から，その数値を引用している統計データにおけるセルを特定するタスクである．つまり，数値とその数値が含まれる文書，被引用統計データが与えられ，数値と一致するセルを特定する．

本研究ではセル特定問題をアドホック検索タスクとして定式化する．つまり，数値と数値を含む文書をクエリ，統計データのセルを文書とみなして，統計データのセルをランキングする．

3.1 定式化

本節では，セル特定問題をアドホック検索問題として定式化する．

入力として与えられたクエリ q と検索対象文書の集合 D に対して，本検索タスクは D に含まれる文書をランク付けする問題であり，出力としてランキング $(d_1, d_2, \ldots, d_k)$ を返す．ただし，各 i について $d_i \in D$ である．

本問題におけるクエリ $q \in Q$ は検証対象の数値と数値を含む文書の組である．ただし，Q は任意のクエリの集合である．

本論文における検索対象文書 $d \in D$ は組 (c, T) で表される．c, T はそれぞれセルと表を表しており，$c \in T, T \in \mathcal{T}$ の関係にある．ただし，$\mathcal{T}$ は統計データを表す．

本論文において統計データとして扱うファイル形式は，スプレッドシート形式 (Excel) もしくは CSV 形式のファイルのみとする．また，本論文においては，統計データ $\mathcal{T}$ を表の集合 $\mathcal{T} = \{T_1, \ldots, T_{|\mathcal{T}|}\}$ として定義する．ここで，スプレッドシート形式のファイルについては，それぞれのシートが統計データに含まれる表 $T_i \in \mathcal{T}$ に対応しているとみなし，CSV 形式のファイルについては，必ず $|\mathcal{T}| = 1$，つまり 1 つの表しか含まれていないとする．

本論文ではこのランキング問題を，クエリ $q \in Q$ と検索対象文書の統計データのセル $d \in D$ を入力として，スコアを出力するスコア関数 $s : Q \times D \to \mathbb{R}$ を設計する問題とみなす．つまり，スコア関数 s の出力するスコアの降順に結果を並べることで最終的なランキングを得る．

4 WikiStatCells データセット

本節ではセル特定問題のためのデータセット WikiStatCells の構築方法について説明した後，構築したデータセットの統計情報について説明する．

4.1 構築手順

本節では Wikipedia 記事と，記事が引用する統計データを用いて WikiStatCells データセットを構築する方法について説明する．

表形式データに関するデータセット作成においては，表全体をアノテータに提示し，アノテーション作業を行ってもらうという形式が一般的である．例えば，表質問応答のデータセットである WikiTableQuestions [10] データセットの作成においては，アノテータにまず Wikipedia に含まれる表を提示し，そこからアノテータは表から回答可能な質問を自由に作成してもらうという手順を取っている．

しかしながら，本論文の扱う表形式データである統計データは，一般の表形式データセットと比較して，行数や列数が大きいという違いが存在する [20]．表形式データセットにおいては Web 上の表が用いられることが多く，後述する通り Web 上の表を用いたセル特定データセットである BriQ [7] においては列数や行数が中央値でともに 5 であるなど，表のサイズが小さいことが多い．しかしながら本研究の扱う統計データは，英語版データセットにおける列数と行数の中央値がそれぞれ 13 と 172 であり，既存のデータセットと比較して表のサイズが大きい．

このような状況において，統計データに対するセル特定データセット作成のために，検証対象の数値とその周辺の文章に合わせて統計データをアノテータに対してそのまま提示すると，アノテータは数値に対応するセルを数千から数万のセルの中から探すこととなる．その結果，1 つのデータを作成するためのアノテーションの作業時間と費用が大きくなり，大規模なデータセットを作成することが困難になってしまうという問題がある．

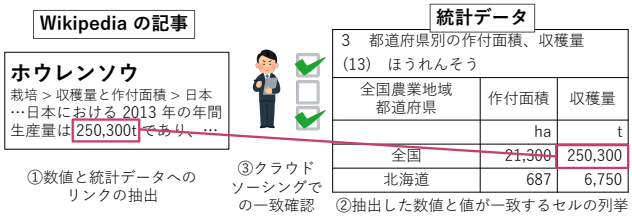

図1　データセット構築手順

2018 年（平成 30 年）のシーベリーの生産量は **5** トンで、栽培面積は 4 ヘクタールでした。

図2　アノテーション対象の数値が含まれる文章の例

そこで本研究では，事前にアノテーションの候補となるセルを絞り込むことによって，アノテーションにかかる時間が少なくなるようにした．セルを絞り込むことにより，アノテータは絞り込まれたセルの行名や列名，表の説明のみを確認するだけで良い．よって，表全体をそのまま提示した場合と比較して，アノテータは膨大なセル数を持つ表の中から対応するセルを特定する必要がなくなり，効率的にアノテーション作業を行うことができるようになると考えられる．

作成手順を図 1 に示す．まず，Wikipedia 記事に含まれる統計データへのリンクを抽出し，そのリンクが含まれる記事のセクションから，引用されている数値の候補を抽出した．次に，リンク先の各統計データに対して，統計データに含まれるセルのうち，抽出した数値と値が一致するセルを機械的に全て列挙した．ただし，抽出された数値と一致するセルの数が 10 個より大きい数値については，アノテーション対象に含むとアノテーション対象のセルの総数が膨大になってしまうため，アノテーション対象から除外した．最後にクラウドソーシングを用いて，列挙したセルの値と抽出した数値が意味的に一致する値であるかを，セルの列名や行名，表の説明などの情報から人手で判定を行った．

クラウドソーシングによるアノテーションのより詳細な設定について説明する．クラウドソーシングに関しては，日本語データセットでは Lancers*1 を，英語版データセットでは Amazon Mechanical Turk*2 を用いた．また，抽出された数値と絞り込まれたセルの組に対して，日本語と英語でともに 3 人ずつのアノテータを割り当てた．

クラウドソーシングにおけるアノテーション対象のデータの例を説明する．アノテーション対象の数値を含む文章の例を図 2 に示す．文章においては，どの数値がアノテーション対象となる数

*1 https://https://www.lancers.jp (2021/12/01 閲覧)
*2 https://www.mturk.com (2021/12/01 閲覧)

	A	B	C	D
2	品目名	栽培面積	収穫量	出荷量
3				
4		ha	t	t
16	シーベリー	4	5	5
17	ブラックベリー	5	5	3

図3　アノテーション対象のセルの例 (意味的に一致する例)

	A	B	C	D
2	品目名	栽培面積	収穫量	出荷量
3				
4		ha	t	t
16	シーベリー	4	5	5
17	ブラックベリー	5	5	3

図4　アノテーション対象のセルの例 (意味的に一致しない例)

表1　アノテーション結果

	日本語	英語
データ数 (一致判定されたセル数)	178	1,268
アノテーション対象セル数	955	6,068
一致判定されたセルの割合	20.2%	26.5%
アノテータ間一致率 (Fleiss's κ)	70.0%	86.2%
Wikipedia 記事数	77	560
統計データ数	26	154
セルあたりのアノテーション時間	1分5秒	1分41秒

値かがハイライトされている．また，上記の数値に対応するアノテーション対象の統計データの例を図 3 に示す．この図の例におけるセル C16 は「シーベリーの収穫量が 5 トンである」ことを表すセルであるため，このセルは文章中の数値と意味的に一致するセルである．一方で，図 4 は文章中の数値と意味的に一致しないセルの例である．これらのセルは文章中のハイライトされた数値と同じ値を持つものの，文脈的に一致しない．例えば C17 のセルは「ブラックベリーの収穫量が 5 トンである」ことを示すセルであり，品目名が異なる．

4.2　アノテーション結果

アノテーション結果を表 1 に示す．日本語データセットにおいては，77 個の Wikipedia 記事と 26 個の統計データから，178 個の統計データを引用する数値と対応する統計データのセルのペアを得ることができた．一方で英語データセットにおいては，560 個の Wikipedia 記事と 154 個の統計データから，1,268 個の統計データを引用する数値と対応する統計データのセルのペアを得ることができた．また，アノテータ間一致率 (Fleiss's κ) は日本語と英語でそれぞれ 70.0%, 86.2% となった．

統計データと Web 上の表を比較するために，作成した統計データに関するデータセットである WikiStatCells (日本語，英語) と既存の Web 上の表のデータセットを比較した．既存の Web 上の表のデータセットとして，セル特定データセットである

表2　データセットに含まれる表に関する比較．値は各データセットにおける中央値を表す．

	日本語	英語	BriQ [7]	WTQ [10]
行数	70	172	5	15
列数	22	13	5	6
数値セル率	89.6%	66.6%	50.0%	20.7%

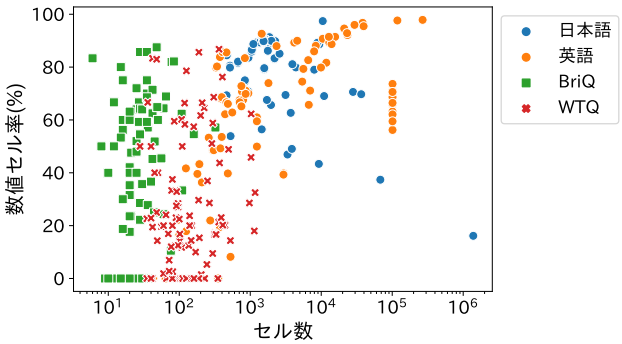

図5　表に含まれるセル数と数値セル率の散布図

BriQ [7] と表質問応答のデータセットである WikiTableQuestions (WTQ) [10] を用いた．結果を表 2 に示す．この表から，作成した WikiStatCells データセット (日本語，英語) は，既存の Web 上の表に関するデータセットと比較して，行数や列数が大きく，かつ数値セル率が高い．

またこれらのデータセットに関して，表に含まれるセル数と表の数値セル率の散布図を図 5 に示す．ただし，表の数が 100 を超えるデータセットはその中から 100 個をランダムにサンプリングしてプロットした．この図において WikiStatCells データセット (日本語，英語) は右上側に多くの点が集まっている．つまり WikiStatCells データセットはセル数が多く，かつ数値セル率が高い表が多い．一方で BriQ や WTQ は左側に多くの点が集まっており，セル数が少ない表が多く，作成したデータセットとは表のサイズに関する違いが存在する．

5　実験

本節では前節で構築したデータセットに対して，ベースライン手法を用いて比較を行う．まずは実験設定について説明した後，実験結果について説明する．

5.1　実験設定

本節では実験設定について説明する．

本実験では 4 つの手法を比較する．具体的には，基本的な検索手法である BM25，クエリ尤度モデル + ディリクレ平滑化 (QLD)，Sequential Dependence Model (SDM) の 3 つに加えて，表質問応答手法である Row-Column Intersection Model (RCI) [5] を比較する． BM25，QLD，SDM は Anserini [16][*3] を用いて実装し，これらの検索手法の持つパラメータは Anserini のデフォ

[*3] http://anserini.io (2021/05/19 閲覧)

表3　実験結果

クエリ	検索手法	日本語		英語	
		MRR	nDCG@10	MRR	nDCG@10
周辺単語のみ	BM25	0.137	0.194	0.100	0.107
	QLD	0.175	0.246	0.110	0.118
	SDM	0.175	0.248	0.109	0.118
	RCI	-	-	0.164	0.184
タイトル+周辺単語	BM25	0.187	0.259	0.195	0.213
	QLD	0.218	0.291	0.294	0.327
	SDM	0.207	0.285	0.295	0.331
	RCI	-	-	0.223	0.247

ルトのものを使用する．

RCI については，HuggingFace Transformers [14] を用いて実装し，WikiTableQuestions データセットで Fine-tuning されたモデルを利用する．ただし，RCI については日本語モデルが存在せず日本語データセットへの適用が難しいため，英語データセットのみに適用する．また，RCI へ適用する際には表のヘッダの特定が必要となるため，本実験ではルールベースでヘッダを抽出する．具体的には，値が空でないセルが 2 個以上含まれる行のうち，最も先頭の行を表のヘッダとみなすこととする．

また，クエリとして用いる情報を変化させた場合について比較する．具体的には，既存の類似する研究 [7, 9] において用いられている，クエリ周辺の情報のみを用いた場合と，クエリ周辺の情報に加えてタイトルを用いた場合を比較する．

クエリ周辺の情報としては数値を含む文を用いることが考えられるが，本データセットにおいては検証対象となる数値が同じ文に複数個含まれる場合が存在する．このような場合，検証対象の数値が異なる場合であってもクエリが同一となり，ランキングの結果も同一となるため，セル特定の性能が低くなる可能性が考えられる．例えば，図 2 の文において「5 トン」と「4 ヘクタール」がともに検証対象の数値であった場合，これらの数値ではクエリが同一の文となる．そこで本実験では検証対象の数値ごとに異なるクエリとするため，クエリ周辺の情報として数値の前後 k 単語ずつ ($k = 10$) を用いることとする． k を変えた場合の性能の変化に関する検証については今後の課題とする．

評価指標としては，アドホック検索においてよく用いられる評価指標である，平均逆数順位（Mean Reciprocal Rank, MRR）と nDCG@10 を用いる．

5.2　結果

本節では実験結果について説明する．実験結果を表 3 に示す．

クエリとして，数値周辺の単語のみを用いた場合と，数値周辺の単語に加えてタイトルも用いた場合を比較すると，全ての手法において，数値周辺の単語に加えてタイトルの情報を用いた場合のほうが評価指標が向上した．この結果から，既存研究の問題設定とは異なり，本論文のセル特定においては，数値周辺の情報に加えてタイトルなどの情報を利用する必要があることが示された．

また，基本的な検索手法 (BM25, QLD, SDM) の結果と表質問応答手法 (RCI) の結果を比較すると，周辺単語のみをクエリとして用いた場合は表質問応答手法が基本的な検索手法を上回る結果となった．一方で，周辺単語に加えてタイトルもクエリとして用いた場合は，BM25 に対しては表質問応答手法が上回る結果となった一方で，QLD と SDM に対しては表質問応答は下回る結果となった．この結果から，本実験で用いた表質問応答手法が質問応答データセットに対して Fine-tuning されていることも考慮すると，表質問応答手法はセル特定問題において強力なベースライン手法となりうることが示された．

6　まとめ

本論文では文書中の数値の真偽を検証するために，数値が参照する統計データのセルを自動的に特定する問題を提案した．我々はこの統計データのセル特定問題をアドホック検索問題と見なし，数値と数値を含む文書の組をクエリ，統計データのセルを検索対象の文書として問題を定式化した．またセル特定問題のためのデータセットとして，Wikipedia と統計データへのリンクを用いて日本語と英語の 2 つの言語からなるデータセット WikiStatCells を作成した．さらに本問題の難しさを検証するために，作成したデータセットを用いて実験を行った．その結果，クエリとしては数値周辺の情報に加えてタイトルの情報を用いた方が検索性能が向上することが判明した．また，ニューラル言語モデルを用いた表質問応答手法がセル特定において強力なベースラインとなりうる可能性が示された．

今後の課題としては，RCI モデルを fine-tuning した場合について検証すること，日本語データセットに対して表質問応答手法を適用することなどに取り組む予定である．

謝辞

本研究はJSPS科研費 18H03244，21H03554，JST さきがけ JPMJPR1853，および JST 次世代研究者挑戦的研究プログラム JPMJSP2124 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Ebrahim Bagheri and Feras N. Al-Obeidat. A latent model for ad hoc table retrieval. In *Proceedings of the 42nd European Conference on IR Research*, pages 86–93. Springer, 2020.

[2] Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D. Davison. Leveraging schema labels to enhance dataset search. In *Proceedings of the 42nd European Conference on IR Research, Part I*, pages 267–280. Springer, 2020.

[3] Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D. Davison. Table search using a deep contextualized language model. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 589–598. ACM, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.

[5] Michael R. Glass, Mustafa Canim, Alfio Gliozzo, Saneem A. Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avi Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. Capturing row and column semantics in transformer based question answering over tables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1212–1224. Association for Computational Linguistics, 2021.

[6] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333. Association for Computational Linguistics, 2020.

[7] Yusra Ibrahim, Mirek Riedewald, Gerhard Weikum, and Demetrios Zeinalipour-Yazti. Bridging quantities in tables and text. In *35th IEEE International Conference on Data Engineering*, pages 1010–1021. IEEE, 2019.

[8] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. A test collection for ad-hoc dataset retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2450–2456. ACM, 2021.

[9] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. Facilitating document reading by linking text and tables. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 423–434. ACM, 2018.

[10] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 1470–1480. The Association for Computer Linguistics, 2015.

[11] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Canim. Ad hoc table retrieval using intrinsic and extrinsic similarities. In *Proceedings of the Web Conference 2020*, pages 2479–2485. ACM / IW3C2, 2020.

[12] Roee Shraga, Haggai Roitman, Guy Feigenblat, and Mustafa Canim. Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1399–1408. ACM, 2020.

[13] Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro A. Szekely. Retrieving complex tables with multi-granular graph representation learning. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*, pages 1472–1482. ACM, 2021.

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 38–45. Association for Computational Linguistics, 2020.

[15] An Yan and Nicholas M. Weber. Mining open government data used in scientific research. In *Proceedings of the 13th International Conference on Information*, pages 303–313. Springer, 2018.

[16] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using lucene. *ACM Journal of Data and Information Quality*, 10(4):16:1–16:20, 2018.

[17] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426. Association for Computational Linguistics, 2020.

[18] Shuo Zhang and Krisztian Balog. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web,*, pages 1553–1562. ACM, 2018.

[19] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103, 2017.

[20] 中野優, 加藤誠. クエリと文書のフィールドを考慮した被引用統計データの検索. 情報処理学会論文誌データベース（*TOD*）, 14(4):49–60, 2021.

[21] 岡本卓, 宮森恒. 被検索文書の絞り込みと補強，クエリ拡張に基づく統計データ向けアドホック検索. 情報処理学会論文誌データベース（*TOD*）, 14(4):36–48, 2021.