被引用統計データのセル特定手法の検討

中野　優† 加藤　誠††

† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{yunakano,mpkato}@acm.org

あらまし　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手法について検討する．我々はまず行数や列数の多い統計データにも適用可能な表質問応答手法を用いてセルを特定する手法を提案する．加えて統計データにおけるセル特定特有の課題を解決するために，1. 文書中の複数の数値を区別するために特殊トークンを導入し，2. 学習データのラベル不均衡を解消するために負例のアンダーサンプリングを行うことを提案する．さらに既存のセル特定データセットを用いて実験を行い，提案手法が比較手法と比べて高い性能となることを示すとともに，提案した 2 つの要素の両方が性能の改善に寄与していることを明らかにした．

キーワード　ドメイン指向アプリケーション（統計データ），検索モデル（言語モデル），質問応答，データ検索，表質問応答

1 はじめに

主張に対してその根拠を示すことは主張の正しさを示す上で重要である．例えば論文において何らかの主張する際には，その主張を支持する根拠がなければ，その主張の正しさを示したことにはならない．その際に統計データを引用して主張を行うことがよく行われており，論文においても近年増加傾向にある [30]．統計データは専門家によって作成された高品質なデータであり，このような信頼性のあるデータを用いて主張の正しさを示すことは重要である．

しかしながら統計データを理解して利用することは容易ではない．多くの統計データは表形式であるが，統計データは行数・列数が非常に多く，また多数の表が含まれることが多い．そのため人手で統計データを読み解き利用することは非常に時間がかかるという問題がある．

このような状況に対して，与えられたテキストと統計データを自動的に紐付ける統計データ検索の研究が行われてきた [6, 19, 36, 38]．これらの研究は事実に対してその根拠として統計データの表を示したものであると言えるが，表の内容は利用者自身が読み解く必要があり，統計データの利活用は未だに時間がかかると考えられる．これに対して，テキストと統計データのセルを直接対応付ける**セル特定問題**に取り組んでいる研究も存在する [4, 16]．しかしながらこれらの研究では，テキストとセルの単語の厳密一致に基いてテキストとセルを対応付けている．そのため利用者がその統計データで用いられる単語を知らなければ単語のミスマッチの問題が発生し，良い精度でテキストとセルを対応付けることは難しい．

そこで本研究ではセル特定問題のためのニューラル言語モデルを用いた新たな手法を検討する．セル特定問題に類似する問題として表質問応答問題が存在する [12, 17, 22, 33]．表質問応答の多くの手法においては質問文と表全体をニューラル言語モデルに入力し，質問に対する答えを抽出・生成する．しかしながら，既存の表質問応答で想定されている表の行数や列数は統計データと比較すると小さい．そのため入力長の制限があるニューラル言語モデルにおいて，既存の表質問応答の手法をそのまま用いることは難しい．

そこで本研究では表全体ではなく，行ごと・列ごとの情報を利用する表質問応答手法 [12] をベースとしたセル特定手法を提案する．この手法においてはまずテキストと表を入力として，行と列のそれぞれについてテキストに適合する確率を計算する．その後，テキストに対する行の適合確率と列の適合確率の積を取ることでテキストとセルの適合確率を計算する．さらにこの手法に対して統計データに対するセル特定問題に特有の問題を考慮し，2 つの改良を行う．1 つ目は特殊トークンを用いた特定対象の明確化である．表質問応答で与えれる質問文とは異なり，セル特定問題で与えられるテキストは複数の数値が含まれ，それぞれが異なるセルに対応付けられる．そこで複数の数値を区別するために，現在着目している数値であることを表す特殊なトークンを用いて数値を囲むことで，より正確にテキストとセルとの対応付けを行えるようにする．

2 つ目は学習データにおける負例のアンダーサンプリングである．統計データは既存の表質問応答で用いられる表よりも行数や列数が多いことが知られている [37]．そのため既存の表質問応答手法と同様に，正解となる行や列を正例とし，統計データ中のそれ以外の行や列を負例として学習すると，学習時間が長くなってしまう．その上，作成された学習データは正例と比較して負例が極端に多いクラス不均衡なデータであり，学習したモデルの性能が低くなる可能性がある．これらの問題を解決するために負例のアンダーサンプリング行うことで，訓練コストを軽減させ，かつ性能を向上させる．

さらに提案した手法の性能を検証するための検証実験を行った．この実験においては，Wikipedia の記事が引用する統計データを用いて作成されたセル特定データセット WikiStatCells [37]

を利用して性能を検証した．その結果提案手法は nDCG@1 において，BM25 などのアドホック検索と比較して約 3.2 倍，表質問応答手法 [12] と比較して約 1.5 倍の性能となることが判明した．さらに Ablation Study を実施し，提案した 2 つの改良の両方が性能向上に寄与していることを示した．

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に対して，ニューラル言語モデルを用いた新たな手法を提案した．

（2） 既存の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い，提案手法の優位性を示した．

本論文の構成は次の通りである．第 2 節ではセル特定問題に関連する研究について説明する．第 3 節でセル特定問題の問題設定について説明した後，第 4 節において提案手法を説明する．第 5 節で検証実験の設定について説明し，第 6 節で実験結果を説明する．第 7 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に関連する問題設定を持つ研究として，2.1 節で表事実検証の研究，2.2 節で文とセルを対応付ける研究，2.3 節で表質問応答の研究を紹介する．その後，関連する研究として 2.4 節で表検索の研究，2.5 節で統計データ検索の研究を紹介する．

2.1 表事実検証

セル特定に関連するタスクとして，本節では表事実検証に関する研究を紹介する．表事実検証（表含意関係認識）とは，文章と表が与えられたときに，表が文章を含意するか否かの 2 値（もしくは無関係を加えた 3 値）に分類するタスクである [5]．表事実検証には 2 種類の手法が存在する．1 つ目は与えられた文章に対して，SQL などの表において実行可能なプログラムを生成する手法である [23, 32, 35]．2 つ目は BERT [9] などのニューラル言語モデルに対して文章と表の関係を学習するために，事前学習とファインチューニングを行う手法である [10, 21]．

しかしながらこれらの手法は単に含意するか否かのみを出力するだけであり，モデルが表のどの部分をもとに答えを判断したが不明であるため，解釈性に欠けるという課題がある．この課題を解決するために，含意するか否かの判断に用いたセルを提示するタスクに取り組む研究も存在する．SemEval-2021 Task 9 においては，表事実検証タスクに加えて証拠セル選択タスクが提案されている [27]．Gupta ら [13, 15] は表が文章を含意するかの根拠となる行を検出する根拠行選択タスクを提案し，データセットを作成している．これらのタスクは文章とセルや行を対応付けているという点でセル特定タスクと類似している．一方で表 1 に示すとおり，既存の表事実検証のデータセットに含まれる表は，統計データと比較して行や列の数が少ない．そのため，多くの手法は表全体がニューラル言語モデルに入力できることを前提としており，そのまま適用することが困難である．

2.2 文とセルを対応付ける研究

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 [18, 20]．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 [20]．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 [18]．

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．

2.3 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 [12, 17, 22, 33]．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 [12]．これまでの多くの表質問応答の研究においては，答えを計算するために SQL などの論理形式を出力する手法が提案されてきた [22, 33]．しかしながらこれらの手法は，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．この課題に対して，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 [12, 17]．これらの手法は，BERT [9] などのニューラル言語モデルの事前学習と弱教師ラベルに基づくによるファインチューニングにより，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，論理形式を経ずに質問に回答することを可能にしている．

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，論理形式を経由し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，セル特定問題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することで，文章に対応するセルを特定できる．しかしながら，セル特定問題と表質問応答は次の 3 つの点で異なる: (1) 本論文の入力は質問文ではなく平叙文である，(2) 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，(3) 本論文が扱う統計データは表質問応答が扱う Web 上の表

表 1　表事実検証とセル特定のデータセットの表の大きさに関する比較．列数と行数はともに中央値.

	TabFact [5]	SEM-TAB-FACTS [27]	InfoTabS [14]	WikiStatCells（英語） [36]
表の出典	Wikipedia の表	論文の表	Wikipedia の INFOBOX	統計データ
行数	12	7	9	172
列数	6	5	2	13

と比較してサイズが大きい.

2.4 表 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, 7, 24–26, 34]．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 [34]．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 [24]，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 [25]，BERT を用いる手法 [7]，グラフニューラルネットワークを用いた手法 [26] 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，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 [37].

2.5 統計データ検索

テキストから統計データを対象として検索を行う研究を説明する [6, 19, 36, 38]．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 [19]．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 [6]．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 [38]．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 [36]．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.

また，与えられたクエリから統計データのセルを対応付ける研究も行われている [2, 4, 16]．Heflin らはセル中心インデックスを提案し，クエリからセルを検索するシステムのためのアーキテクチャとユーザインタフェースを提案した [16]．Cao らは統計データ検索において，適合する統計データに合わせて適合するセルを表示するために，セルをスコア付けする方法を提案した [4]．Balalau らは Cao らの手法に基づいて，統計データとそのセルを検索するためのシステムを構築した [2]．これらの研究はテキストからセルを検索し対応付けているという点で本研究と関連しているが，次の 2 点で異なる．1 つ目は既存研究は単語の厳密一致に基づくセル検索手法を採用している点である．もしユーザが統計データ中で利用されている語彙に詳しくない場合，単語の厳密一致に基づく手法は検索性能が低くなると考えられる．これは統計データの内容に詳しくないことが予想される非専門家がユーザである場合に問題となる．2 つ目はこれらの研究は統計データ（表形式データ）の検索の補助としてセルを提示しており，提示されたセルがクエリに適合するかを評価していない点である．そのため，これらの研究において提案されたセル検索の手法がそれ以外の手法と比較して良い手法であるかは明らかではない.

3 問 題 設 定

本節ではセル特定タスクの問題設定と定式化について説明する．セル特定タスクとは，文書中の数値が与えられたときに，数値に対応する統計データのセルを特定する問題である．セル特定問題の定式化を行う前にまず統計データについて定義し，その後セル特定問題の定式化を説明する.

a） 表と統計データ

まず表と統計データについてそれぞれ定義する．T_R 行 T_C 列を持つ表 T を，ヘッダ $H_T = (h_1, \ldots, h_{T_C})$ とセル値 $V_T = (v_{i,j})$ $(1 \leqq i \leqq T_R, 1 \leqq j \leqq T_C)$ からなる組 $T = (H_T, V_T)$ と定義する．ここで，ヘッダ h_i やセル値 $v_{i,j}$ はテキストデータであり，単語列とみなせる．これらは整数や実数などの数値と解釈できる場合もあるが，本論文ではテキストデータとして扱う．このとき統計データ $\mathcal{T}$ を表の集合 $\mathcal{T} = \{T_1, \ldots, T_{|\mathcal{T}|}\}$ と定義する．このように定義される統計データとして，Excel などのスプレッドシート形式のデータや CSV 形式のデータが該当する.

b） 数値を含む文書

本節では数値と数値を含む文書をそれぞれ定義する．数値 v^q を単語列と定義する．このとき数値 v^q を含む文書 d について，数値 v^q を連続部分列として含む単語列 $d = (w_1, \ldots, w_n)$ であると定義する．つまり，ある $1 \leqq i \leqq j \leqq n$ について $v^q = (w_i, \ldots, w_j)$ であり，文書 d は 3 つの列 $\mathrm{ctxt}_l, v^q, \mathrm{ctxt}_r$ を連結した列として表せる．ただし

$\mathrm{ctxt}_l = (w_1, \ldots, w_{i-1}), \mathrm{ctxt}_r = (w_{j+1}, \ldots, w_n)$ とする．

数値 v^q は文書 d において引用されている統計データの集合 $\mathrm{cited}(d)$ に含まれる統計データ $\mathcal{T} \in \mathrm{cited}(d)$ に含まれるいずれかのセルに対応しているとする．ただし数値 v^q は値の丸めや誤引用などの影響により，数値 v^q と値が一致するセルが統計データ $\mathcal{T}$ において存在するとは限らないものとする．

c） セル特定問題

本節ではセル特定タスクをアドホック検索問題として定義する．つまり．クエリ $q \in Q$ 検索対象の統計データの集合 D に対して，本検索タスクは D に含まれる統計データのセルをランク付けする問題であり，出力としてランキング $(v_1, v_2, \ldots, v_K)$ を返す．ただし，各 k $(1 \leqq k \leqq K)$ について v_k はある統計データ $\mathcal{T} \in D$ に含まれるセルとする．

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

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 $\mathcal{T} \in D$ に含まれるセル v を入力として，スコアを出力するスコア関数 $s : Q \times \mathcal{V} \to \mathbb{R}$ 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．ただし $\mathcal{V}$ は統計データの集合 D に含まれる全ての統計データの表のセルの集合であるとする．

4 提案手法

本節ではまずクエリに対するセルのスコアを得るためのモデルについて説明する．その後，クエリと統計データの入力表現について説明し，クエリの入力表現における特殊なトークンを用いたクエリの数値の明確化についても説明する．最後に，学習データのクラス不均衡問題に対処するための負例のアンダーサンプリングについて説明する．

4.1 モデル

本節ではセルの確率を得るためのモデルについて説明する．

まず，既存の表質問応答手法とその問題点について説明する．2.3 節で説明した通り，論理形式を経由しない表質問応答手法はセル特定可能であると考えられる．例えば TaPaS [17] は質問文と表全体を BERT [9] などのニューラル言語モデルに入力し，どのセルを質問の回答に用いるかを計算する．そのため，このセルの確率が高い順にセルを並べ替えることで，クエリに対するセルのランキングを作成できる．

ここで，既存の表質問応答で用いられているデータセットに含まれる表は行数や列数が統計データよりも少ないことが知られている [37]．そのため表全体を入力したとしても，ニューラル言語モデルが持つ入力長制限がモデルの性能に影響を与えることは少ないと考えられる．一方で統計データのようにの表の行数や列数が多い場合においては，TaPaS のような表全体を入力とするモデルを用いると，ニューラル言語モデルの入力長制限がモデルの性能に影響を与える可能性が大きくなると考えられる．

そこで本研究では，行と列ごとにクエリとの適合確率を判定する行・列交差モデル（Row-Column Intersection，以下 RCI）[12] を採用する．RCI は表全体ではなく行ごと・列ごとにその内容をニューラル言語モデルへ入力する．そのため，たとえ行数や列数が大きい場合であっても，表全体を入力する場合よりも入力の際に失われる情報が少なくなるという利点を持つ．

RCI におけるクエリに対するセルの適合確率を計算する方法を説明する．まずクエリ q と行 row_i の組を Cross-Encoder モデルの入力として，クエリが行に適合する確率 p_{row_i} を得る．また，同様の方法でクエリ q が列 col_j に適合する確率 p_{col_j} を得る．つまり，式にすると以下の通りである．

$$p_{\mathrm{row}_i} = \mathrm{softmax}(\mathrm{MLP}(\mathrm{TF}_{\mathrm{row}}(\tau_{q,\mathrm{row}_i})_{[\mathrm{CLS}]}))_1 \quad (1)$$

$$p_{\mathrm{col}_j} = \mathrm{softmax}(\mathrm{MLP}(\mathrm{TF}_{\mathrm{col}}(\tau_{q,\mathrm{col}_j})_{[\mathrm{CLS}]}))_1 \quad (2)$$

ただし，$\tau_{q,\mathrm{row}_i}, \tau_{q,\mathrm{col}_j}$ はそれぞれ次節で説明するクエリと行の入力表現，クエリと列の入力表現を表す．また，$\mathrm{TF}_{\mathrm{row}}(\cdot), \mathrm{TF}_{\mathrm{col}}(\cdot)$ はそれぞれ行と列のためのトランスフォーマーモデル（ニューラル言語モデル）を表し，本論文ではこれらのモデルをそれぞれ行モデル，列モデルと呼ぶ．$\mathrm{TF}_{\cdot}(\cdot)_{[\mathrm{CLS}]}$ は Transformer の最終層のベクトルのうち [CLS] トークンのベクトルを表す．

最終的に行と列の確率の積により，クエリがセルに適合する確率を計算する．つまりクエリ q とセル $v_{i,j}$ に対して，スコア関数 $s(q, v_{i,j})$ を以下で定義する．

$$s(q, v_{i,j}) = p_{\mathrm{row}_i} \cdot p_{\mathrm{col}_j} \quad (3)$$

このスコアを用いてクエリ q に対するセルのランキングを作成する．

4.2 クエリと統計データの入力表現

本節ではクエリと統計データの入力表現について説明する．

前節で述べた通り，本研究で用いる RCI は Cross-Encoder モデルであるため，クエリと行の組，クエリと列の組をそれぞれ行モデル $\mathrm{TF}_{\mathrm{row}}(\cdot)$ と列モデル $\mathrm{TF}_{\mathrm{col}}(\cdot)$ の入力とする．つまり，行モデルと列モデルの入力表現 $\tau_{q,\mathrm{row}_i}, \tau_{q,\mathrm{col}_j}$ は以下のように書ける．

$$\tau_{q,\mathrm{row}_i} = [\mathrm{CLS}]\tau_q[\mathrm{SEP}]\tau_{\mathrm{row}_i}[\mathrm{SEP}] \quad (4)$$

$$\tau_{q,\mathrm{col}_j} = [\mathrm{CLS}]\tau_q[\mathrm{SEP}]\tau_{\mathrm{col}_j}[\mathrm{SEP}] \quad (5)$$

ただし，$\tau_q, \tau_{\mathrm{row}_i}, \tau_{\mathrm{col}_j}$ はそれぞれクエリ q，行 row_i，列 col_j の入力表現とする．

a） クエリの入力表現

3 節で述べたとおり，本研究のセル特定問題におけるクエリは検証対象の数値 v^q と数値を含む文書 d の組である．単純には文書 d をそのまま入力に用いることが考えられるが，文書 d は一般に長いため，ニューラル言語モデルの入力長制限が問題となる．

そこで文書の一部のみを用いることで，この問題を解決する．

具体的には，文書のタイトルなどのメタデータと，数値を含む文などの数値の周辺情報を入力として用いることとする．

さらに，クエリとなる数値を明確にするために，周辺情報のうちクエリとなる数値 v^q をクエリであることを表す特殊トークン [NUM], [/NUM] で囲む．このように，入力表現中の特に着目すべき対象を明確化するために特殊トークンを用いる方法はエンティティ検索 [28] や関係分類 [3]，共参照解析 [29] などのタスクで用いられている．セル特定においても同じ文に複数の検証対象の数値が存在する場合があり，どの数値が検証対象の数値かを明確にすることはセル特定の性能につながると考えられる．

よって，クエリ $q = (v^q, d)$ の入力表現 τ_q は以下のようになる．

$$\tau_q = \mathrm{meta}_d \text{ [/META] } \mathrm{ctxt}_l \text{ [NUM] } v^q \text{ [/NUM] } \mathrm{ctxt}_r \quad (6)$$

ただし，meta_d は文書 d のメタデータを表し，$\mathrm{ctxt}_l, \mathrm{ctxt}_r$ は数値 v^q のそれぞれ左側，右側の周辺情報を表す．また，[/META] はメタデータと数値の周辺情報との区切りを表す追加の特殊トークンを表す．

b） 統計データの入力表現

行の入力表現 row_i と列の入力表現 col_j は RCI と同様の入力表現を用いる．具体的には以下のように，ヘッダとセルの値を用いた入力表現を用いる．

$$\tau_{\mathrm{row}_i} = h_1 : v_{i,1} \mid h_2 : v_{i,2} \mid \cdots \mid h_{T_C} : v_{i,T_C} \mid \quad (7)$$

$$\tau_{\mathrm{col}_j} = h_j \ : \ v_{1,j} \mid v_{2,j} \mid \cdots \mid v_{T_R,j} \mid \quad (8)$$

4.3 負例のアンダーサンプリング

本節では学習データにおける負例のアンダーサンプリングについて説明する．

まず，RCI における学習データの作成とその問題点を説明する．RCI においては，クエリの数値に対応するセルが属する行（列）をクエリに対する正例となる行（列）とし，それ以外を負例とみなして学習データを作成する．しかしながら，本研究が対象とする統計データは既存の表質問応答や表事実検証のデータセットに含まれる表よりも，表の行数や列数が多い [37]．そのため正例に加えてすべての負例を学習に用いた場合，学習コストが大きくなるという問題がある．

また学習データにおけるクラス不均衡により，モデルの性能が低下する可能性が懸念される．既存の表事実検証データセットである InfoTabS [14] データセットにおいては，正例 1 つあたり負例は約 6 個である [13]．一方で，セル特定データセット WikiStatCells [37] においては表 2 に示す通り，行モデルでは正例 1 つあたり負例は 1000 個以上となる．このように，正例に対して負例が極端に多いクラス不均衡データにおいて，そのまま学習を行うとモデルの性能が低くなるという問題がある．

これら 2 つの問題を解決するために，負例のアンダーサンプリングを行う．具体的には 1 つのクエリに対して最大 n_{neg} 個の負例を全体からサンプリングし，正例とサンプリングした負例を用いて学習を行う．これにより，学習の高速化と性能の向上が期待できる．

表 2　5 交差検証における Fold ごとのデータセットの統計情報（Fold ごとの平均）．負例の括弧内は訓練時の負例のアンダーサンプリング後の値を表す．

	クエリ数	正例	負例
行モデル	254	279	296,595 (29,453)
列モデル	254	318	17,489 (9,607)

5 実験設定

5.1 データセット

本節では評価実験に利用するデータセットについて説明する．

本論文では中野と加藤が作成したデータセットである WikiStatCells [37] を利用して評価実験を行う．このデータセットは Wikipedia とその記事が引用する統計データを利用して作成されたセル特定タスクのためのデータセットである．つまり Wikipedia 記事中の数値に対して，その記事が引用する統計データ中のセルを人手で対応付けられたデータセットとなっている．

本実験ではこのデータセットのうち英語版データセットを用いて実験を行う．英語版データセットにおいては英語版 Wikipedia の 560 個の記事に含まれる 1,268 個の数値に対して対応する統計データのセルが付与されている．統計データとしてはアメリカの国勢調査局とイギリス国家統計局の 154 個の統計データを対象としており，ファイル形式としては Excel ファイルと CSV ファイルを対象としている．

5.2 統計データの前処理

提案手法の統計データの入力表現においては列ヘッダが必要となる．しかしながら，本実験で利用する統計データは Excel ファイルや CSV ファイルであり，ファイル中のどの部分がヘッダであるかを判定する必要がある．

そこで本実験においては，既存手法を用いて Excel ファイルや CSV ファイルに含まれる表のどの部分がヘッダであるかを抽出する．具体的には，Excel ファイルと CSV ファイルのそれぞれについて，Ghasemi-Gol らの手法 [11] と Christodoulakis らの手法 [8] を用いてヘッダの抽出を行う．

5.3 データセット分割

本実験では 5 交差検証で性能を検証する．表 2 に各 Fold ごとの統計情報を示す．

またデータセット分割を行う際に，ある 2 つのクエリの正解セルが同じ統計データに含まれる場合には，その 2 つのクエリが同じテストデータのグループに含まれるように交差検証の分割を行う．つまりこの分割においては，テストデータで正解となるセルを含む統計データが，学習時に含まれないように分割が行われる．

このような分割を行う理由は，本データセットの特有の性質を考慮するためである．本データセットにおいては，ある 2 つのクエリの正解セルが同じ統計データに含まれる場合において，

正解となる 2 つのセルの行もしくは列が同じになりやすいという性質がある．もし前段落で述べた分割を行わない場合，列モデルや行モデルは訓練データにおいて正例となる行や列を覚えておき，テスト時においてもその行や列を上位とすれば性能が良くなってしまう．つまり，このように学習された列モデルや行モデルはクエリの意味を理解せずに統計データのみから適合・不適合の判定を行っている．そのため，たとえこのようなモデルにおいて性能が高くなったとしても，真にセル特定の問題が解けているとは言い難いと考えられる．このような問題を回避するために，テストデータで現れる統計データが学習時に出現しないように分割を行う．

5.4 ベースライン手法

本実験では 2 種類のベースライン手法と比較を行う．1 種類目は単語の厳密一致によるアドホック検索手法である．具体的には，BM25，クエリ尤度モデル + ディリクレ平滑化 (QLD)，Sequential Dependence Model (SDM) の 3 つの手法を用いる．実装としては Anserini [31] を利用し，統計データの各セルを 1 つの文書としてインデックスする．セルをインデックスする際には，そのセルを含む列のヘッダとそのセルを含む行のヘッダからなる文字列をセルに対応する文書とする．また，各検索手法のパラメータは Anserini のデフォルトのパラメータを利用する．

2 種類目は表質問応答手法である．具体的には，提案手法のベースとなった RCI [12] と比較を行う．この手法は提案手法から 4.2 節で説明した対象の数値の明確化と 4.3 節で説明した負例のアンダーサンプリングを行わない場合の手法とみなせる．実装については次節の提案手法の実装と同様に行う．

5.5 提案手法の実装

提案手法のモデルの学習について説明する．提案手法の行モデル，列モデルの学習においては，ベースとなる RCI [12] の設定にしたがって学習を行う．4.3 節で説明したサンプリングする負例の数は，クエリ 1 つにつき $n_{\mathrm{neg}} = 128$ 個とする．4.2 節で説明した追加の特殊トークン [NUM], [/NUM], [/META] は事前学習時は存在しないトークンである．そのため提案手法の学習時（ファインチューニング時）において，事前学習済みモデルの埋め込み層に対して，追加する特殊トークンに対応する埋め込みベクトルを追加した上で学習を行う必要がある．本実験ではランダムに初期化したベクトルを用いて，特殊トークンに対応する埋め込みベクトルをモデルの埋め込み層に追加した．

クエリとなる数値を含む文書の設定について説明する．またクエリの文書表現においては，文書のメタデータ meta_d として文書 d のページタイトルと，数値 v^q が属するセクションとそのセクションの全ての上位セクションのセクションタイトルを用いる．数値 v^q の周辺情報 $\mathrm{ctxt}_l, \mathrm{ctxt}_r$ としては，数値 v^q を含む文を用いる．

5.6 評価指標

本実験では 3 つの評価指標を用いる．具体的には，アドホック検索においてよく用いられている平均逆数順位（Mean Reciprocal Rank，以下 MRR），nDCG@1，nDCG@10 を用いる．実装としては PyNTCIREVAL[1] を用い，nDCG@{1,10} は Microsoft バージョンの MSnDCG を用いて計算する．

表 3 実験結果．提案手法の結果については独立に 5 回学習を行った結果の平均と標本標準偏差を表している．

	MRR	nDCG@1	nDCG@10
BM25	0.311	0.192	0.352
QLD	0.312	0.181	0.355
SDM	0.311	0.175	0.360
RCI [12]	0.532	0.405	0.568
提案手法	0.698 (±0.012)	0.618 (±0.017)	0.715 (±0.017)

表 4 Ablation Study の結果．括弧内は提案手法の結果との差分を表している．

	MRR		nDCG@1	
提案手法	0.654		0.554	
特殊トークンなし	0.564	(-0.090)	0.435	(-0.118)
アンダーサンプリングなし	0.606	(-0.048)	0.521	(-0.032)
両方なし	0.532	(-0.122)	0.405	(-0.148)

6 実験結果

本節では WikiStatCells データセットでの実験結果について説明する．加えて，Ablation Study として提案手法の要素ごとの検証の結果と，列モデルと行モデルのそれぞれのモデルごとでの結果について説明する．

6.1 実験結果

実験結果を表 3 に示す．この表より提案手法は，BM25 などの単語の厳密一致に基づくアドホック検索や表質問応答手法である RCI と比較して，全ての評価指標において良い性能となった．具体的には，比較手法の中で最も良い性能となった RCI と比較して，MRR で 0.16 ポイント以上，nDCG@1 で 0.21 ポイント以上，nDCG@10 で 0.14 ポイント以上の改善が見られた．

6.2 Ablation Study

本節ではセル特定の性質を考慮した改良に関する Ablation Study の設定とその結果について述べる．

Ablation Study として，4 節で説明した提案手法について，手法に含まれる特定の要素を除いた場合の手法について性能の検証を行った．具体的には，以下の 3 つの手法について実験を行った．

（1） 特殊トークンなし: 4.2 節で説明した，クエリとなる数値の明確化のために用いた特殊トークンを使わない場合の手法である．

1：https://github.com/mpkato/pyNTCIREVAL（2022/01/10 閲覧）

表 5　行モデルと列モデルの結果．提案手法の結果については独立に 5 回学習を行った結果の平均と標本標準偏差を表している．

		MRR	nDCG@1	nDCG@10
行モデル	RCI [12]	0.659	0.553	0.701
	提案手法	0.749 (±0.009)	0.640 (±0.014)	0.782 (±0.009)
列モデル	RCI [12]	0.732	0.603	0.780
	提案手法	0.843 (±0.014)	0.788 (±0.020)	0.864 (±0.012)

（2） アンダーサンプリングなし: 4.3 節で説明した，学習データの負例のアンダーサンプリングを行わない場合の手法である．

（3） 両方なし: 上記 2 つの両方を行わない場合の手法であり，比較手法である RCI [12] に相当する手法である．

また，負例のアンダーサンプリングについては，1 クエリあたり 100 個の負例を使用して実験を行った．

Ablation Study の結果を表 4 に示す．この表の結果から，特殊トークンなし，アンダーサンプリングなし，両方なしのいずれの場合においても，提案手法と比較して性能が低下した．よって，クエリとなる数値の明確化のための特殊トークンと，学習データのラベル不均衡のための負例のアンダーサンプリングの両方の工夫が性能の改善に寄与していると考えられる．

6.3 行モデルと列モデル

本節では行モデルと列モデルの結果について説明する．4 節で説明した通り，RCI のアーキテクチャに基づいた提案手法は，行と列のそれぞれについてクエリとの適合確率を計算し，その確率をもとにセルとの適合確率を計算する手法であった．そのため，行のモデル（行モデル）と列のモデル（列モデル）のそれぞれについてランキングを作成し，データセットに対して評価指標を計算することが可能である．

行モデルと列モデルのそれぞれに対して評価指標を計算した結果を表 5 に示す．この表の結果から，行モデルと列モデルの両方において，すべての評価指標で RCI よりも性能が改善していることがわかった．

7 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手法について検討した．我々はまず行数や列数の多い統計データにも適用可能な表質問応答手法を用いてセルを特定する手法を提案した．加えて統計データにおけるセル特定特有の課題を解決するために，1. 文書中の複数の数値を区別するために特殊トークンを導入し，2. 学習データのラベル不均衡を解消するために負例のアンダーサンプリングを行うことを提案した．さらに既存のセル特定データセットを用いて実験を行い，提案手法が比較手法と比べて高い性能となることを示すとともに，提案した 2 つの要素の両方が性能の改善に寄与していることを明らかにした．

今後の課題としては，データやモデルに基づいた負例のアンダーサンプリングの改善を行うことが考えられる．

謝辞　本研究は JSPS 科研費 21H03554，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．

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] Oana Balalau, Simon Ebel, Théo Galizzi, Ioana Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. Statistical claim checking: Statcheck in action. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4798–4802. ACM, 2022.

[3] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Tien Duc Cao, Ioana Manolescu, and Xavier Tannier. Searching for truth in a database of statistics. In *Proceedings of the 21st International Workshop on the Web and Databases, Houston, TX, USA, June 10, 2018*, pages 4:1–4:6. ACM, 2018.

[5] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *Proceedings of 8th International Conference on Learning Representations*, 2020.

[6] 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.

[7] 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.

[8] Christina Christodoulakis, Eric Munson, Moshe Gabel, Angela Demke Brown, and Renée J. Miller. Pytheas: Pattern-based table discovery in CSV files. *Proc. VLDB Endow.*, 13(11):2075–2089, 2020.

[9] 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Julian Eisenschlos, Syrine Krichene, and Thomas Müller. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online, November 2020. Association for Computational Linguistics.

[11] Majid Ghasemi-Gol, Jay Pujara, and Pedro A. Szekely. Tabular cell classification using pre-trained cell embeddings. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 230–239. IEEE, 2019.

[12] Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem 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, Online, June 2021. Association for Computational Linguistics.

[13] Vivek Gupta, Riyaz A. Bhat, Atreya Ghosal, Manish Shrivastava, Maneesh Singh, and Vivek Srikumar. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *Transactions of the Association for Computational Linguistics*, 10:659–679, 2022.

[14] Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online, July 2020. Association for Computational Linguistics.

[15] Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3283, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Jeff Heflin, Brian D. Davison, and Haiyan Jia. Exploring datasets via cell-centric indexing. In *Proceedings of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021*, volume 2950 of *CEUR Workshop Proceedings*, pages 53–60. CEUR-WS.org, 2021.

[17] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian 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, Online, July 2020. Association for Computational Linguistics.

[18] 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.

[19] 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.

[20] 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.

[21] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. TAPEX: table pre-training via learning a neural SQL executor. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[22] 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 (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics.

[23] Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. Learn to combine linguistic and symbolic information for table-based fact verification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5335–5346, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[24] 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.

[25] 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.

[26] 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.

[27] Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online, August 2021. Association for Computational Linguistics.

[28] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online, November 2020. Association for Computational Linguistics.

[29] Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online, July 2020. Association for Computational Linguistics.

[30] 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.

[31] 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.

[32] Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825, Online, November 2020. Association for Computational Linguistics.

[33] 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, Online, July 2020. Association for Computational Linguistics.

[34] 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.

[35] Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. LogicalFactChecker: Leveraging logical operations for fact checking with graph module network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6065, On-

line, July 2020. Association for Computational Linguistics.
[36] 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.
[37] 中野優, 加藤誠. 被引用統計データのセル特定データセットの構築. **日本データベース学会論文誌データドリブンスタディーズ**, 1(1), 2023.
[38] 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.