一般発表 | Track 3: 情報検索・情報推薦・ソーシャルメディア

2026年3月1日(日) 13:00 ~ 15:10 | F会場

## [5F] 統計データと評価データ

座長:金子 邦彦(福山大学)　コメンテータ:天方 大地(大阪大学)

13:00 ~ 13:25

### [5F-01] 統計データを用いた事実確認支援のための統計データ内の関連箇所抽出

*宮﨑 隆豪[1]、宮森 恒[1] (1. 京都産業大学　情報理工学部　宮森研究室)

13:25 ~ 13:50

### [5F-02] 事実確認支援のための言説と統計データ関連箇所との整合性検証

*樫山 和貴[1]、宮森 恒[1] (1. 京都産業大学)

13:50 ~ 14:15

### [5F-03] 統計データ検索における視覚的文書検索の有効性分析

*福岡 啓人[1]、宮森 恒[1] (1. 京都産業大学)

14:15 ~ 14:40

### [5F-04] 日本語検索タスクにおける機械翻訳テストコレクションの妥当性検証

*岩間 悠莉[1]、加藤 誠[2,3] (1. 筑波大学 知識情報・図書館学類、2. 筑波大学 図書館情報メディア系、3. 国立情報学研究所)

14:40 ~ 15:05

### [5F-05] 事前学習済みBERTモデル検索タスクのための評価データセット

*ファム フーロン[1]、三林 亮太[4]、荘司 慶行[5]、加藤 誠[3,6]、山本 岳洋[1]、山本 祐輔[2]、大島 裕明[1] (1. 兵庫県立大学、2. 名古屋市立大学、3. 筑波大学、4. 神戸大学、5. 静岡大学、6. 国立情報学研究所)

# 統計データを用いた事実確認支援のための統計データ内の関連箇所抽出

宮﨑　隆豪†　　宮森　　恒†

† 京都産業大学情報理工学部情報理工学科　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2254711,miya}@cc.kyoto-su.ac.jp

**あらまし**　本稿では，ネット上の言説に対する統計データを用いた事実確認支援のための関連箇所抽出の問題に取り組む．省庁などが提供している統計データは，一定の信頼性が担保された情報と捉えることができるため，それらを用いた事実確認はネット上の言説の真偽判断支援に有用と考えられる．これを実現するには，関連する統計データの検索，統計データ内の関連箇所抽出，ネット上の言説と統計データ内関連箇所との整合性検証の 3 段階の処理が必要となるが，本稿では，この第 2 段階の問題を扱う．統計データ内の関連箇所抽出を実現するには，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットが必要となるが，従来研究ではそのようなデータセットは提供されていない．そこで本稿では，この課題を解決するため，大規模言語モデルを活用したデータセット構築手法を提案する．具体的には，統計データとその内部の関連箇所を入力とし，それに整合する，あるいは，整合しない言説を LLM に生成させることで，大規模かつ質の高いデータセット構築を試みた．実験では，構築したデータセットの妥当性を検証するとともに，このデータセットを活用して関連箇所抽出モデルの性能を評価する．

**キーワード**　事実確認支援，統計データ，関連箇所抽出，ソーシャルメディア，説明文生成

## 1　はじめに

近年，SNS やオンラインプラットフォームは日常における主要な情報源として定着しており，日常的に多種多様な情報がやり取りされている．これらの情報は社会や個人の意思決定に大きな影響を与える一方で，その信頼性が常に保証されているわけではない．特に，誤った情報や悪意のある情報が拡散されることで，社会的混乱を招くリスクが高まっている．例えば，パンデミック時には，誤った医療情報が広がり，正しい予防策の普及や治療への適切な対応が遅れる事例が報告されている．また，政治的な偽情報が選挙結果や政策決定に影響を及ぼす可能性も指摘されている．このような背景から，情報の真偽を正確に判断するための「事実確認支援（fact-checking support）」が社会的に重要な課題となっている．

こうした状況において，信頼性の高い情報源として，政府や公共機関が提供する統計データの重要性が増している．統計データは客観的な数値に基づく情報を提供し，主観や憶測が含まれやすいネット上の言説に対し，客観的かつ確定的な証拠能力を持つ．したがって，統計データを用いた事実に即した検証は，言説の真偽を客観的に判断するための強固な基盤となる．

統計データを活用した事実確認支援を実現するためには，以下の 3 段階の処理が必要と考えられる．まず，事実確認対象の言説に関連した統計データ検索 [1]，次に，検索された統計データ内から言説の事実確認に利用できる関連箇所の特定，最後に，特定した関連箇所と言説との整合性の検証である．例えば，「日本の人口は増加している」という言説に対して，適切な統計データを用いてその真偽を判断するには，統計データ内の総人口に関するセクションを特定し，近年の人口推移を確認した上で言説内容との整合性を検証する必要がある．

本稿では，事実確認支援の第 2 段階である「統計データ内の関連箇所抽出」に焦点を当てる．この処理を効果的に実現するためには，統計データの内容を適切に解釈し，言説に対応する箇所を的確に特定する必要がある．しかし，従来研究では，このようなタスクに必要なデータセットが整備されておらず，適切な関連箇所抽出の実現に向けた基盤が十分ではなかった．また，統計データは数百から数千行に及ぶことも珍しくなく，それらをすべて大規模言語モデルの入力ウィンドウに収めることは，技術的な制限や情報密度の過剰な上昇を招く．これにより，LLM が表の構造を正しく把握できず，存在しない数値を参照したり，行を読み間違えたりするハルシネーションを引き起こし，抽出精度が著しく低下するという課題も存在する [2].

本稿では，これらの課題に対応するため，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを構築し，その有用性を示すことを目的とする．本稿で対象とする統計データには，公的機関が提供する信頼性の高い CSV 形式のデータを用いる．まず前処理として，複雑な構造を持つ CSV データを，解析が容易な 2 次元のデータフレーム形式へと変換する処理を行う．これにより，表形式データ特有の行列構造を保持したまま，計算機での効率的な処理を可能とする．その後，統計データの背景情報を示すメタデータと統計データ内の特定の一行である関連箇所を LLM への入力とし，その関連箇所に基づいた言説を生成することで，統計データ，言説，関連箇所の対応関係が明確なデータセットを構築する．この際，ソーシャルメディアなどで見られる多様な言説の真偽の度合いを反映させるため，生成される言説には真偽の度合いが異なる 4 種類のラベルを付与する．

本稿では，構築したデータセットを用い，LLM による関連箇所抽出の性能を評価するための実験を実施する．前述の通り，

統計データは膨大な数値情報を含むため，一度の推論ですべての情報を精査することは困難であるという課題がある．そこで本稿では，データを 10 行単位のバッチに分割して段階的に絞り込みを行う「2 段階推論システム」と，独自の「関連度スコア」を用いた手法を提案し，抽出精度と処理効率の両面からその有効性を検証する．本稿の主な貢献は以下の通りである．

(1) 言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを構築し，統計データを活用した事実確認支援の基盤を提供する．

(2) 大規模な統計データを効率的に処理するバッチ分割型の 2 段階推論システムと，独自スコアリングによる高精度な関連箇所抽出手法を提案し，統計データ内の関連箇所抽出を実現した．

(3) 構築したデータセットと提案手法を用いた実験を通じて，統計データを用いた事実確認支援における LLM の適用可能性と，本データセットの妥当性を示した．

本稿の構成は以下の通りである．第 2 節では，事実確認支援および表形式データの理解に関する関連研究について述べる．第 3 節では，統計データ内の関連箇所抽出における問題設定を定義し，本稿で提案する 2 段階推論システムと関連度スコアリング手法の詳細を説明する．第 4 節では，提案手法の検証に用いるためのデータセット構築手順とその諸元について述べる．第 5 節では，構築したデータセットを用いた実験結果を示し，提案手法の有効性および今後の課題について考察する．最後に，第 6 節において本稿のまとめと今後の課題について述べる．

## 2 関 連 研 究

### 2.1 関連箇所抽出

機械読解（Machine Reading Comprehension; MRC）は，与えられたテキストから質問に対する回答を抽出するタスクであり，特にスパン抽出（Span Extraction）は，回答を文中の一部分として特定する手法として広く用いられている．代表的なデータセットとして SQuAD [3] があり，文中の開始位置と終了位置を特定する形式での応答が求められる．MRC におけるスパン抽出手法は，BERT [4] や SpanBERT [5] など事前学習モデルを用いることで性能の向上が進んでいる．

しかし，従来の抽出型機械読解は，回答が一つの範囲に限定される単一回答 QA が中心であった．これに対し，DROP [6] などのデータセットでは，複数の範囲における回答抽出が必要な複数回答 QA が新たに追加されており，複数の回答がコンテキストに散在する場合に対応するための研究も進められている．本稿は，言説の根拠となる統計データ内の「関連箇所」を特定するものであり，テキストベースの MRC で培われたスパン抽出技術を，表形式データに適用するための基礎研究として位置づけられる．

### 2.2 表形式データの学習

表や統計データを活用した自然言語処理 (NLP) の分野では，テキスト情報と表データを統合的に扱う手法が注目されている．

TaBERT [7] は，表データの構造を考慮し，BERT を基盤としたエンコーダを用いて，自然言語テキストと表の情報を統合的に学習するモデルである．これにより，表の構造やセル情報を考慮しながら，質問応答 (QA) や検索タスクに応用可能な表現を学習できる．

また, TAPAS [8] も, 表形式データ上の推論に特化した BERT ベースのモデルとして提案されている．TAPAS は，表のキャプションや記事のタイトルなどを質問文の代わりとして用い，Wikipedia の大規模な表形式データで事前学習を行うことで，表のセルを選択したり，集計操作を行ったりするタスクを解くことが可能である．これらの研究は，表形式データの構造を理解し，その内容を自然言語で扱うための基盤技術である．しかし，TaBERT や TAPAS といった既存の表形式データ専用エンコーダを用いた手法は，モデルの入力制限により，本稿が対象とするような数百行を超える大規模な統計データ全体を一度に処理することが困難である．これに対し，本稿では特定の表形式データ用エンコーダは採用せず，汎用的な大規模言語モデルを推論エンジンとして用いる．これにより，専用モデルの追加学習コストを抑えつつ，第 3 節で提案する「2 段階推論システム」を通じて，大規模な表構造の中から必要な情報を柔軟かつピンポイントに特定することが可能となる．

### 2.3 事実確認支援

事実確認支援 (fact-checking support) は，オンライン上の誤情報や悪意のある情報の拡散を防ぐための重要な研究領域である．特に，ソーシャルメディアやニュースサイトにおける情報の信頼性を検証するため，多くの手法が提案されている．代表的なデータセットとして FEVER [9] があり，Wikipedia を基に 18 万件以上の主張と，それに対応する証拠文のペアを提供する．FEVER では，事実確認を「主張が支持されるか，反証されるか，情報不足か」を判定するタスクとして定義し，自然言語推論 (NLI) や検索技術と組み合わせた多くの手法が提案されている．例えば，文書検索と文選択を組み合わせたパイプライン手法 [10] や，グラフ構造を用いて複数の証拠文を統合し推論を行う手法 [11] などが提案され，高い精度を達成している．

また，近年では Transformer ベースのモデルを活用した誤情報検出技術も発展している．DisinfoBERT [12] は，ソーシャルメディア上の誤情報を識別するために設計された BERT ベースのモデルであり，文のコンテキストや言語的特徴を考慮して誤情報を分類している．このような技術は，ニュース記事や SNS の投稿の信頼性を判断する上で有用であり，誤情報対策として実用化が進められている．

本稿は，これらの事実確認の枠組みと共通する課題を扱うが，対象とするデータが異なる点に特徴がある．既存の研究は主にニュース記事や Wikipedia を対象とし，テキストベースのファクトチェックを行うのに対し，本研究では公的機関が提供する信頼性の高い統計データを直接的な根拠とする言説の収集・整理を可能にする．ネット上の書き込みと統計データの関連箇所を特定することで．主張と数値データの整合性を評価し，ファクトチェックを補助する仕組みを構築することを目的としている．

## 3 提案手法

本節では，第 1 節で述べた統計データ内の関連箇所抽出という課題に対し，本研究が提案する 2 段階推論システムおよび関連度スコアリング手法の詳細について述べる．

### 3.1 問題設定

本稿で取り組む統計データ内の関連箇所抽出タスクは，自然言語で記述された言説と対象となる統計データを入力とし，その言説の事実確認を行う上で参照すべき統計データ内の関連箇所を特定することを目的とする．

本タスクの入出力を次のように定義する．言説の集合を $C$，統計データの集合を $D$ とする．言説 $c \in C$ に対し，対象とする統計データ $d \in D$ は，背景情報を示すメタデータ $m$，表の構造を定義するヘッダー $h$，および表内の $n$ 行の行集合 $R = \{r_1, r_2, \ldots, r_n\}$ の組として次のように表される．

$$d = (m, h, R) \tag{1}$$

ある言説 $c$ と統計データ $d$ に対して，各行 $r_i \in R$ が言説の根拠としてどの程度関連しているかを示す関連度スコアを $s_{r_i}$ とする．本タスクの目的は，$c, m, h, r_i$ を入力として $s_{r_i}$ を算出する関数 $f$ を用い，スコア $s_{r_i}$ が最大となる行 $r^*$ を関連箇所として特定することである．

$$s_{r_i} = f(c, m, h, r_i) \tag{2}$$

$$r^* = \arg\max_{r_i \in R} s_{r_i} \tag{3}$$

本稿では，LLM に対して「言説 $c$」「メタデータ $m$」「ヘッダー $h$」「行 $r_i$」の要素を提供し，これらを統合的に解釈させることで関数 $f$ を実現し，関連箇所の特定を行う．なお，本稿では計算機による一貫した管理と言説生成の確実性を担保するため，抽出の最小単位を統計データ内の特定の一行として定義する．詳細なデータセットの構成および各ラベルの定義については第 4 節で述べる．

### 3.2 2 段階推論システムによる抽出プロセス

大規模な統計データから効率的かつ高精度に関連箇所を特定するため，本稿では図 1 に示すような「2 段階推論システム」と「関連度スコアリング」を用いた手法を提案する．具体的な抽出プロセスを以下に定義する．

#### 3.2.1 バッチ分割

まず，入力となる統計データの行集合 $R$ を，$k$ 行（本稿では $k = 10$）ずつの連続する部分集合であるデータバッチ $B = \{b_1, b_2, \ldots, b_m\}$ に分割する．ここで，$m = \lceil |R|/k \rceil$ である．各データバッチ $b_j$ は，LLM が表構造を識別しやすいよう Markdown 形式に変換され，各行には明示的な行番号が付与される．

#### 3.2.2 バッチフィルタリング

第一段階では，言説 $c$，メタデータ $m$，ヘッダー $h$，および各データバッチ $b_j$ を LLM に入力し，当該バッチと言説の関連度を示すバッチスコア $s_{b_j}$ を算出する．

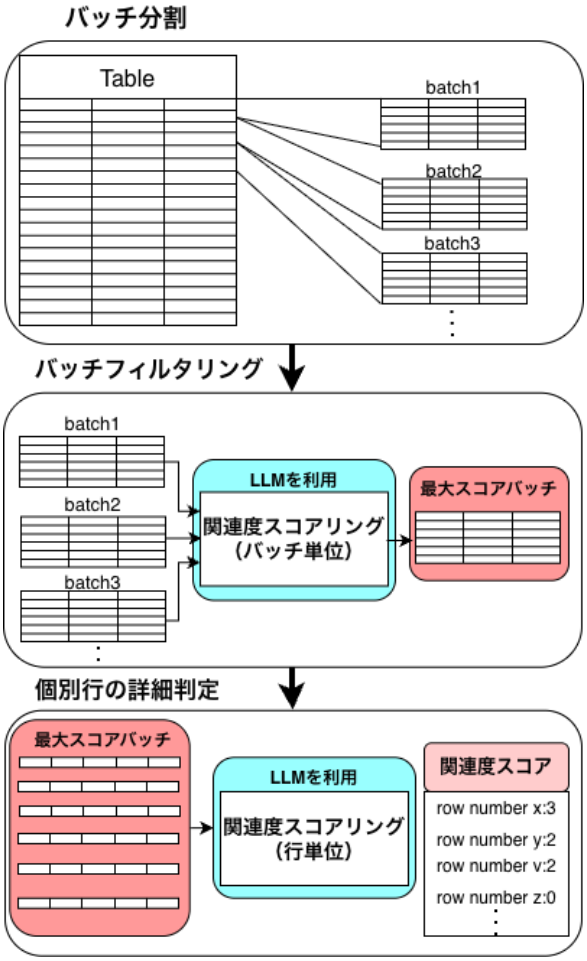



図 1　関連箇所抽出の概要図

$$s_{b_j} = f_{stage1}(c, m, h, b_j) \quad (s_{b_j} \in \{0, 1, 2, 3\}) \tag{4}$$

次に，スコア集合 $\{s_{b_1}, \ldots, s_{b_m}\}$ の中から最大スコア（閾値 $s \geq 1$ ）を獲得したバッチ $b_{target}$ を詳細解析の対象として選別する．

$$b_{target} = \arg\max_{b_j \in B} s_{b_j} \tag{5}$$

全てのバッチのスコアが 0 であった場合は，該当箇所なしと判定し処理を終了する．

#### 3.2.3 個別行の詳細判定

第二段階では，$b_{target}$ に含まれる各行 $r_i \in b_{target}$ に対してより緻密な解析を行う．ここでは，項目名と数値の対応を強調するため，各行を「フィールド名：値」のテキスト形式に整形して提示する．LLM はこれに基づき，各行の関連度スコア $s_{r_i}$ を算出する．

$$s_{r_i} = f_{stage2}(c, m, h, r_i) \quad (s_{r_i} \in \{0, 1, 2, 3\}) \tag{6}$$

最終的に，あらかじめ設定した閾値（$s \geq 2$ ）を満たし，かつ最大スコアを持つ行 $r^*$ を言説の根拠となる関連箇所として特定する．

$$r^* = \arg\max_{r_i \in b_{target}} s_{r_i} \tag{7}$$

表 1　関連度スコアの定義と判断基準

| スコア | 定義 |
|---|---|
| 3 | データ行の情報のみで，言説の主要な要素の真偽が完全に確定する． |
| 2 | データ行の情報が，言説の主要な要素の真偽を部分的に判断できる証拠を提供する（一部の要素は不明だが，重要な手がかりとなる）． |
| 1 | データ行の情報が，言説の背景，文脈，または間接的なヒントとして関連するが，真偽判定には不十分である． |
| 0 | データ行の情報と，言説の主要な要素の間に意味のある接点がない． |

### 3.3　関連度スコアリング

本手法における各段階の推論では，LLM に対して言説とデータの関連性を数値化させる「関連度スコア」を導入する．スコアは 0 から 3 の 4 段階の整数値で定義され，その具体的な判断基準を表 1 に示す．評価にあたっては，主観的な解釈や推測を排除し，提供されたデータ行が言説の真偽を判定するための直接的な根拠となり得るかという観点から判断を行うよう，プロンプトにて指示する．

各段階におけるスコアの活用方法は以下の通りである．第一段階のバッチフィルタリングにおいては，閾値を 1 と設定し，スコア 1 以上のバッチの中から最大スコアを獲得したバッチを次段階の解析対象として選出する．第二段階の詳細推論においては，閾値を 2 と設定し，スコア 2 以上の行を抽出対象とする．このように，段階的に閾値と提示形式を調整することで，膨大なデータの中から真偽判定に寄与する核心的な情報の特定を試みる．

## 4　データセット構築

### 4.1　データセット構築の概要

本稿では，統計データを用いた事実確認支援のための基盤を構築することを目的とし，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットを作成する．この 3 つ組のうち，統計データについては，NTCIR-15 [13] で提供される公的データを利用する．NTCIR-15 の統計データは，多様な分野のデータを網羅し，標準化された形式で提供されていることから，データの信頼性が高く，事実確認支援の基盤として活用するのに適している．一方，ネット上で事実確認支援の対象候補となる言説は，統計データと紐付けられておらず，データセットとして利用可能なデータが不足している．このため，本稿では，NTCIR-15 の統計データを活用し，それに関連する言説を生成するために LLM を用いた言説生成手法を採用する．

本稿では，言説，統計データ，関連箇所の 3 つ組から構成されるデータセットの構築を問題として設定し，その具体的な要件を以下のように定める．

まず，事実確認の対象となる自然言語の「言説」に対し，その根拠となる統計データ内の「関連箇所」が明確に紐付けられている必要がある．本稿では，言説生成の確実性とデータセットの構造的一貫性を保つため，関連箇所を統計データ内の特定の一行として定義する．統計データにおける関連箇所は，本来「単一のセル」「行内の一部の列」「離れた複数行」など多様なパターンが存在するが，本稿で行単位の定義を採用した主な理由は，以下の 2 点に集約される．

まず，統計データを用いた事実確認支援の第 3 段階である整合性検証において，LLM は抽出された特定行のみならず，その周辺行をコンテキストとして参照することで，統計データの時系列的な変化やカテゴリ間の比較といったメタ情報を踏まえた高度な推論を行うことが可能である．したがって，第 2 段階において行単位の特定が実現できれば，整合性判定に必要な情報は十分に保持されると言える．

次に，実用上の柔軟性が挙げられる．バッチ分割による 2 段階推論と関連度スコアリングを組み合わせた本手法では，個別の行に対して独立にスコアリングを行うため，関連箇所が離れた複数行に及ぶ場合でも，各行に対して高い関連度を付与することで，漏れなく第 3 段階の整合性検証へと情報を引き渡すことができる．また，特定のセルに依存する言説であっても，そのセルを含む「行」全体を入力として与えることで，第 3 段階における数値の特定とその意味解釈を同時に行うことが可能となる．

以上の理由から，本稿では計算効率と推論に必要な情報のバランスを考慮し，抽出の最小単位を行として定義する．これにより，どの行の情報と言説が対応しているかを厳密に管理しつつ，後続の処理へ十分な情報量を担保することが可能となる．

次に，現実の事実確認の状況を反映させるため，生成される言説は多様な真偽の度合いを持つ必要がある．ソーシャルメディアなどで見られる言説は，必ずしも完全に正しいものや誤っているものばかりではなく，一部に事実を含むものや，提示された情報だけでは判断できないものも存在する．このような現実の状況を再現するため，本データセットでは統計データに基づいて生成される言説に対し，表 2 に示す 4 種類のラベルを定義し，付与することとする．

これらの要件を満たすデータセットの構築は，手動アノテーションの莫大なコストや専門知識の必要性といった課題を伴う．本稿は，これらの課題に対し，特に LLM を活用したデータセット構築アプローチを提案することで，統計データを用いた事実確認支援のための関連箇所抽出技術の発展に寄与することを目指す．

### 4.2　データセット構築の手順

本節では，統計データ，言説，および関連箇所の 3 つ組から構成されるデータセットの構築手順について述べる．本手法では，人手による大規模なアノテーションコストを削減しつつ，質の高いデータを効率的に生成するため，LLM を活用した自動構築プロセスを採用する．具体的な手順は以下の 2 段階から

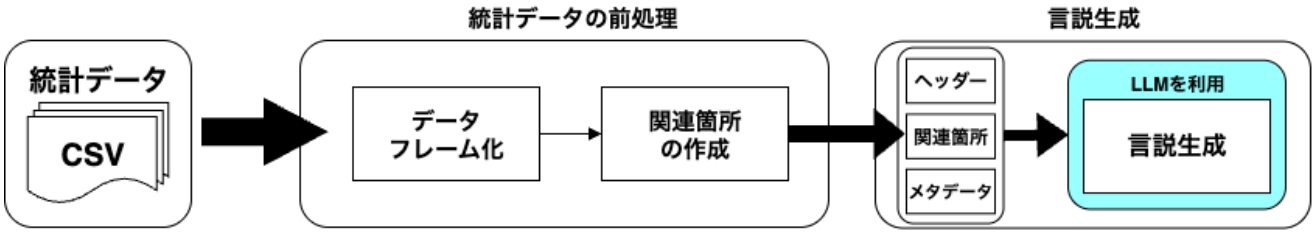



図 2　データセット構築の概要図

表 2　言説に付与するラベルの定義

| ラベル | 定義 |
|---|---|
| True（真） | 言説の内容が，対応する統計データによって完全に支持される． |
| False（偽） | 言説の内容が，対応する統計データによって明確に反証される． |
| Partially True（一部真） | 言説の一部は統計データによって支持されるが，他の一部は裏付けられない，あるいは矛盾する内容を含む． |
| Undeterminable（判別不能） | 言説は統計データに関連する内容ではあるが，そのデータのみでは真偽を判断することができない． |

なる（図 2）．

(1) 統計データの前処理とデータフレーム化

まず，NTCIR-15 から提供される CSV 形式の統計データを取得し，前処理を行う．複雑な構造を持つ CSV データを，LLM が論理的に解釈しやすい 2 次元のデータフレーム形式に変換する．この際，行列ヘッダの階層構造を整理し，各行が独立した意味を持つ単位として抽出可能な状態に整える．

(2) LLM による言説の自動生成とラベル付与

次に，データフレーム化された統計データからランダムに抽出した「関連箇所（特定の 1 行）」と，表の構成を定義する「ヘッダー」，およびその統計データの背景情報である「メタデータ」を LLM に入力する．ヘッダー情報を併せて入力することで，LLM は関連箇所に含まれる数値情報の意味を正確に把握し，文脈に即した言説を生成することが可能となる．LLM には，与えられた数値情報に基づき，表 1 で定義した 4 種類のラベル（True, False, Partially True, Undeterminable）のいずれかに合致する自然言語の言説を生成させる．これにより，関連箇所と言説の対応関係が厳密に担保されたデータセットを構築する．

### 4.3　統計データの前処理

本節では，データセット構築の基盤となる「統計データの前処理」の詳細について述べる．本工程の目的は，多様な形式で存在する公的統計データを，一貫した構造を持つデータフレーム形式へと変換することにある．

まず，元の CSV データに対して，不要な連続した数値列や特定の文字列 (改行文字，カンマなど)，および空白行の削除といった前処理を適用し，データのクリーンアップを行う．次に，クリーンアップされたデータの中から，表の数値領域の左上部分の位置を特定し，その情報を用いて行ヘッダおよび列ヘッダを抽出する．一部の統計データに見られる多階層構造のヘッダに対しても，直前までの行・列の状態を保持・比較することで適切に結合し，単一のヘッダ行として展開する処理を適用する．最後に，抽出されたヘッダ情報と数値領域を対応付け，最終的なデータフレームを構築する．前述の通り，本稿では関連箇所の最小単位を「行」として定義しており，このデータフレームの各行が言説生成の根拠として抽出される．このように構造化された形式で入力を行うことで，LLM は数値情報の意味を正確に把握し，文脈に即した言説生成が可能となる．なお，本工程で構築されたデータフレームは，第 5 節で述べる提案手法の関連箇所抽出タスクにおいて，LLM が論理構造をより解釈しやすいよう Markdown 形式へと変換した上で入力に利用される．

### 4.4　言 説 生 成

本節では，データセット構築の中核となる，大規模言語モデルを用いた言説生成手法について詳述する．言説生成には，高い推論性能と多岐にわたるタスクへの適応性を持つ OpenAI 社の GPT-4o [14] を利用する．

具体的な生成プロセスとして，LLM には，統計データの背景情報を提供するメタデータ，統計データのヘッダー，および，統計データ内の数値領域（ヘッダー行を除くデータ行）からランダムに抽出された「特定の 1 行（関連箇所）」を入力し，これらに基づいた主張文を生成させる．行をランダムに選択することで，統計データ内の特定の項目に依存しない，網羅的かつ客観的な言説の収集を可能にしている．

この際，関連箇所の該当行のみを入力するのではなく，その周辺前後 5 行を含めたコンテキストを提示する工夫を施している．これにより，LLM は関連箇所の数値が統計表全体の中でどのような位置付けにあるのかを正確に把握することが可能となり，表の構造を誤解することなく，より自然かつ妥当性の高い言説の生成が期待できる．

また，本稿では，4.1 節で定義した 4 種類のラベルに基づき，プロンプトを通じてそれぞれ異なる要件を含めることで，特定の性質を持つ言説を生成させる．これにより，単に事実と整合する言説を生成するだけでなく，虚偽の言説や判別不能な言説など，現実の事実確認シナリオに対応した多様な言説を体系的に作成する．

### 4.5　データセットの諸元

本手法を用いて構築したデータセットの具体的な統計量を表 3 に示す．本データセットは，NTCIR-15 から選定した 200 件の

統計データに基づき構築された．各統計データに対して 4 種類の真偽ラベル（True, False, Partially True, Undeterminable）の言説をそれぞれ 3 件ずつ，合計 12 件生成しており，データセット全体では 2,400 件の言説で構成される．

対象とした統計データの規模は，最小 2 行から最大 7,479 行と極めて幅広く，平均行数は 1,142.4 行に達する．表 3 の行数分布に示す通り，500 行を超える大規模なデータが全体の半数以上（54.5%）を占めており，多くの統計データにおいて LLM の最大トークン制限を超えるサイズが含まれている．この結果は，本稿で提案するバッチ分割型の 2 段階推論システムの必要性を裏付けるものである．

表 3　構築したデータセットの統計量および内訳

| 項目 | 件数・数値 | 割合 |
|---|---|---|
| **統計データの行数分布** | | |
| 1 ～ 100 行 | 42 件 | 21.0% |
| 101 ～ 500 行 | 49 件 | 24.5% |
| 501 ～ 1,000 行 | 25 件 | 12.5% |
| 1,001 ～ 2,000 行 | 25 件 | 12.5% |
| 2,001 行以上 | 59 件 | 29.5% |
| **統計データ全体** | | |
| 対象ファイル総数 | 200 件 | 100.0% |
| 累積総データ行数 | 228,474 行 | - |
| 平均行数 / ファイル | 1,142.4 行 | - |
| **生成言説の内訳（ラベル別）** | | |
| True / False / Partial / Undet. | 各 600 件 | 各 25.0% |
| 合計言説数 ($N$) | 2,400 件 | 100.0% |

## 5 実　　験

### 5.1 実験目的

本稿では，第 4 節で構築したデータセットの妥当性を検証するとともに，統計データ内の関連箇所抽出における提案手法の性能を評価するための実験を実施する．

第一の目的は，提案手法の構成要素である「バッチ分割による 2 段階推論」および「関連度スコアリング」の有効性を検証することである．具体的には，これらの要素の有無を組み合わせた複数の手法を比較評価するアブレーション実験を行い，従来手法である一括入力方式に対する優位性を，推論の成功率と抽出精度の両面から明らかにする．

第二の目的は，提案手法を用いた際の言説の性質やモデルの違いによる挙動の差を分析することである．ここでは提案手法に固定した上で，言説のラベル別に関連箇所抽出の精度を算出する．あわせて，バックエンドに用いる LLM として商用モデルとオープンソースモデルを比較し，ラベルごとの抽出難易度やモデル間の性能差について詳細な知見を得ることを目的とする．

### 5.2 実験設定

実験に用いるモデル，比較手法，データセット，および評価指標について説明する．

#### 5.2.1 使用するモデル

推論エンジンには，高い推論能力を持つ商用モデルである OpenAI 社の GPT-4o，およびオープンソースモデルである Alibaba Cloud 社の Qwen2.5-14B-Instruct [15] を採用した．Qwen2.5-14B は，パラメータ数が比較的軽量でありながら高い日本語処理能力を保持しており，ローカル環境やプライベートクラウドでの運用を想定した実用的なモデルとしての性能を検証するために採用した．これにより，クラウド型の巨大モデルと、運用コストやデータ秘匿性に優れた軽量モデルとの間における性能差を明らかにする．

#### 5.2.2 比較手法

本稿では，提案する「2 段階推論」および「関連度スコアリング」の有効性を明らかにするため，これらの構成要素の有無を組み合わせた 4 つの手法を設定し，比較評価を行う．

まず，バッチ分割を行わず，統計データの全行およびメタデータを単一のプロンプトに集約して入力する「一括方式」として，スコアリングを行わずに関連行の直接的な抽出のみを指示する「一括方式（スコアなし）」と，各行に対するスコアリング結果に基づき抽出を行う「一括方式（スコアあり）」を設定する．これらの一括方式においては，提案手法との公平な比較を期すため，入力データは提案手法と同様の Markdown 形式に変換し，各行に行番号を付与して提示する．なお，一括方式は統計データが長大である場合に LLM のトークン制限を超過する恐れがあるため，本実験ではこのような入力制限に対する手法の堅牢性についても評価の対象とする．

これらに対し，第 3.2 節で述べたバッチ分割および 2 段階推論を採用した手法として，各段階での判定を関連の有無のみの二値判定とする「2 段階方式（スコアなし）」と，これに各段階での関連度スコアリングを統合した「提案手法（2 段階＋スコアあり）」を比較に用いる．

#### 5.2.3 実験データセット

本実験では，第 4 節で構築したデータセットの中から，評価用として以下の条件に基づき抽出したサブセットを利用する．

まず，構成要素の有無によるアブレーション実験においては，各ラベルから 30 件ずつ抽出した合計 120 件の言説を用いる．また，GPT-4o と Qwen2.5-14B を用いたモデル別の詳細比較実験においては，各ラベルから 50 件ずつ抽出した合計 200 件の言説を用いる．

データセットの選定にあたっては，実験結果が特定のデータ規模に依存することを防ぐため，統計データの行数分布が元のデータセット全体の平均的な分布（表 3）と整合するよう配慮した．この条件を満たす範囲内でランダムに抽出を行うことで，大規模なデータから小規模なデータまでを網羅し，かつ客観的な評価が可能なサブセットを構築した．

### 5.3 評価方法と評価指標

本実験では，提案手法による関連箇所抽出の性能を定量的に評価するため，各言説に対し正解となる関連箇所は常に 1 行であると定義し，以下の手順と指標を用いて評価を実施する．

**5.3.1** 評価の手順

提案手法については，プログラムの処理ステップに合わせ，Stage 1（バッチフィルタリング）および Stage 2（個別行の詳細判定）の 2 段階で評価を行う．まず，Stage 1 ではデータバッチの中から正解行を含むバッチを正しく選別できているかを検証する．次に，Stage 2 において，選別されたバッチ内から最終的に正解行を特定できたかを検証する．

一方，一括入力方式については，単一の推論結果において正解行が最大スコアを獲得しているかを検証する．なお，適合率，再現率，および F1 スコアの算出にあたっては，正常に推論を完了した試行のみを評価対象とし，実行に失敗した試行は集計から除外する．

**5.3.2** 評 価 指 標

本稿では，実験の目的に応じて以下の指標を使い分けて評価を行う．構成要素の有無によるアブレーション実験では，手法の堅牢性と基本性能を測るため，実行成功率，適合率，再現率，F1 スコア，および平均実行時間を用いる．ラベル別・モデル別の詳細分析では，提案手法の内部動作を詳細に評価するため，実行成功率に代わりバッチ特定成功率を導入し，これに適合率，再現率，F1 スコア，平均実行時間を加えて評価を行う．各指標の定義は以下の通りである．

**1. 実行成功率**：入力データが LLM のトークン制限などに抵触せず，正常に推論を完了した割合を示す．大規模な統計データに対する各手法の「堅牢性」を評価する指標である．

**2. バッチ特定成功率（BSR; Batch identification Success Rate）**：Stage 1 において，正解行を含むバッチが最大スコアを獲得し，次段階へ正しく引き継がれた割合である．提案手法の有効性を内部的に評価する指標として用いる．

**4. 適合率 (Precision)・再現率 (Recall)・F1 スコア (F1-score)**：モデルが「関連あり」と判定した行を対象に算出する．適合率は抽出された行のうち正解が占める割合，再現率は正解行を漏らさず抽出できた割合，F1 スコアはその調和平均である．

**5. 平均実行時間**：1 言説あたりの処理に要した時間の平均値であり，手法やモデル間の実用的な処理効率を比較するために用いる．

## 5.4 実 験 結 果

本節では，5.3 節で定義した各指標に基づき，関連箇所抽出の評価結果を述べる．まず，アブレーション実験を通じて各構成要素の有効性を検証し，次に複数の LLM を用いたラベル別の詳細な性能比較を行う．

**5.4.1** アブレーション実験による構成要素の評価

表 4 に，GPT-4o を用いた提案手法および各比較手法の性能評価結果を示す．

まず，実行成功率に着目すると，一括方式（スコアなし・あり）がいずれも 70%台に留まっているのに対し，2 段階方式を採用した手法はいずれも 100.0%の成功率を記録した．大規模な統計データに対しても，バッチ分割による 2 段階推論を用いることで，トークン制限を回避し安定して推論を実行できることが確認された．実行時間については，一括方式と比較して大幅に増加する傾向にあるものの，データの規模に関わらず安定した処理が可能であるという利点がある．

次に，抽出精度（F1 スコア）を比較すると，一括方式（スコアあり）が 0.817 と高い値を示している一方で，提案手法（2 段階＋スコアあり）は 0.784 であった．ただし，一括方式の結果は実行に成功した 70%のデータのみを対象とした集計値であるのに対し，提案手法は失敗事例を含む全データ（$N = 120$）を完遂した上での数値である．また，スコアリングの有無による影響を見ると，それぞれスコアありの手法がスコアなしの手法の F1 スコアを上回る結果となった．

**5.4.2** モデル別のラベル別抽出精度

各ラベル 50 件（合計 $N = 200$）のデータセットを用い，提案手法を GPT-4o および Qwen2.5-14B に適用した際の性能比較を表 5 に示す．

全体平均の性能を確認すると，バッチ特定成功率（BSR）については，GPT-4o が 0.969，Qwen2.5 が 0.943 といずれも高い数値を示した．これは，バックエンドのモデルの種類に関わらず，提案手法の第一段階（バッチフィルタリング）が，正解を含むバッチを極めて高い確率で次段階へ引き継げていることを示している．

一方で，ラベル別の抽出精度を確認すると，モデルによって精度に顕著なばらつきが確認された．GPT-4o では「True」ラベルにおいて F1 スコアが 0.860 と最も高かったのに対し，Qwen2.5 では「Partially True」が 0.640 と最も高く，「True」は 0.542 であった．特に Qwen2.5 の「True」ラベルでは，再現率（Recall）が 0.930 と高い一方で適合率（Precision）が 0.382 と低く，他のラベルと比較して適合率の低下が顕著であった．

また，「Undeterminable」ラベルにおいても，両モデルで異なる傾向が見られた．GPT-4o は再現率（0.692）と適合率（0.729）が近い値となったのに対し，Qwen2.5 は適合率が 0.810 と全項目中で最高値を示した一方で，再現率は 0.462 と低い値に留まった．

平均実行時間については，GPT-4o（374.51 秒）に対し，Qwen2.5（1738.39 秒）は約 4.6 倍の時間を要しており，商用モデルとオープンソースモデルの間で処理効率に顕著な差が見られた．

## 5.5 考　　察

本実験の結果に基づき，提案手法の有効性とデータセットの妥当性，および今後の課題について考察する．

第一に，提案手法の妥当性について考察する．表 4 の結果より，提案手法は実行成功率 100.0%を達成した．一括入力方式（スコアあり）の F1 スコア（0.817）は数値上は提案手法（0.784）を上回っているが，これは全データの約 3 割におよぶエラー事例を除外した，成功事例のみの平均値である点に留意する必要がある．本手法のようにデータをバッチ分割して段階的に絞り込むアプローチは，一度に処理する情報密度を最適化し，長大なデータに対しても安定した出力を可能にした．実行時間の大幅な増加という課題はあるものの，データの規模に関

表 4　提案手法および各比較手法の性能比較（$N = 120$）

| 比較手法 | 実行成功率 | Precision | Recall | F1 スコア | 平均時間 (s) |
|---|---|---|---|---|---|
| 一括方式（スコアなし） | 71.0% | 0.709 | 0.833 | 0.766 | **6.78** |
| 一括方式（スコアあり） | 70.0% | 0.764 | **0.876** | **0.817** | 9.15 |
| 2 段階方式（スコアなし） | **100.0%** | **0.813** | 0.732 | 0.771 | 262.93 |
| 提案手法（2 段階＋スコアあり） | **100.0%** | 0.798 | 0.771 | 0.784 | 316.73 |

表 5　モデル別のラベル別抽出精度の詳細比較（$N = 200$）

| ラベル | モデル | BSR | 適合率 | 再現率 | F1 スコア | 平均時間 (s) |
|---|---|---|---|---|---|---|
| True | GPT-4o | **1.000** | **0.761** | **0.989** | **0.860** | **390.72** |
| | Qwen2.5 | 0.930 | 0.382 | 0.930 | 0.542 | 1306.85 |
| False | GPT-4o | 0.921 | **0.569** | **0.842** | **0.679** | **375.43** |
| | Qwen2.5 | **0.923** | 0.489 | 0.718 | 0.582 | 2268.03 |
| Partially True | GPT-4o | 0.955 | **0.617** | **0.864** | **0.720** | **396.46** |
| | Qwen2.5 | **0.971** | 0.511 | 0.857 | 0.640 | 1498.66 |
| Undeterminable | GPT-4o | **1.000** | 0.729 | **0.692** | **0.710** | **335.43** |
| | Qwen2.5 | 0.949 | **0.810** | 0.462 | 0.588 | 1880.00 |
| 全体平均 | GPT-4o | **0.969** | **0.669** | **0.847** | **0.742** | **374.51** |
| | Qwen2.5 | 0.943 | 0.548 | 0.742 | 0.588 | 1738.39 |

わらず処理を完遂できる「堅牢性」は，実務における事実確認支援システムとして不可欠な要素であると言える．

第二に，抽出精度における指標間の特性差について考察する．表 5 の結果より，全体として適合率よりも再現率が高い傾向が見られた．これは，モデルが正解行を確実に含めるために，正解の周辺行や類似した属性を持つ行に対しても高いスコアを付与する「過剰検知」が発生したためと考えられる．特に，輸出入統計のように類似した数値や項目名が並ぶデータにおいて，LLM が直接的な根拠だけでなく比較対象となる背景情報まで「関連あり」と広義に捉える傾向が確認された．

第三に，ラベル別の精度差とモデルごとの特性について考察する．本実験では「False」や「Undeterminable」の F1 スコアが低迷する傾向が見られたが，その要因はモデルによって異なると推察される．GPT-4o においては，言説と統計データ内の数値が一致しない場合に「関連なし」と判断する傾向が強く，数値の表層的な一致を優先したことが精度低下を招いたと考えられる．一方，Qwen2.5 においては，「True」ラベルで再現率が極めて高い（0.930）一方で適合率が極端に低い（0.382）という特徴が見られた．これは Qwen2.5 が数値の一致・不一致に関わらず，特定のキーワード等に基づいて過剰に関連箇所を抽出する傾向があることを示唆している．このように，モデルの種類によって，真偽ラベルに応じた抽出の挙動に明確な差異が存在することが確認された．

第四に，実用上の運用形態と今後の課題について述べる．本実験の結果，一括入力が可能な小規模なデータに対しては一括入力方式が精度面で優位であり，大規模なデータに対しては 2 段階推論が堅牢性の面で不可欠であることが明らかになった．したがって，実システムにおいては，入力される統計データの行数やトークン量に応じてこれら 2 つの手法を動的に切り替えるハイブリッド型の抽出アルゴリズムが有効であると考えられる．

今後は，どの程度のデータ規模を閾値として手法を切り替えるべきかの最適化に加え，過剰検知を抑制するためのプロンプトエンジニアリングや，関連度スコアに対する適切な閾値の設定が必要である．また，数値が不一致である場合でも，それが「否定の根拠」であることを LLM に正しく認識させるための推論プロセスの改善も，事実確認支援の精度向上において不可欠な課題であると言える．

## 6 ま と め

本稿では，ネット上の言説に対する統計データを用いた事実確認支援の 3 段階のうち，第 2 段階である「統計データ内の関連箇所抽出」に焦点を当て，その実現のために必要となる書き込みテキスト，統計データ，関連箇所の 3 つ組から構成されるデータセット構築を試みた．データセット構築においては，統計データ内の特定の一行を関連箇所とし，当該箇所の数値情報と統計データの背景情報であるメタデータを LLM に入力することで言説を生成した．その際，生成される言説には実際のソーシャルメディアで見られるような多様な真偽の度合いを反映した 4 種類のラベルを付与し，現実の事実確認の状況を再現する基盤を構築した．

本データセットを用いた関連箇所抽出実験では，データのバッチ化と独自の「関連度スコア」に基づく段階的な絞り込みを行う「2 段階推論システム」を提案し，その有効性を検証した．実験の結果，一括入力方式ではトークン制限により実行成功率が 70%台に留まったのに対し，提案手法は 100.0%を達成し，大規模な統計データに対する堅牢性を示した．また，関連度スコアリングを用いた判定により，提案手法の第一段階におけるバッチ特定成功率（BSR）においても高い精度を記録し，

膨大なデータの中から参照すべき箇所を効率的に絞り込めることを示した．

一方で，適合率が再現率を下回る傾向が見られたことから，類似した数値が並ぶデータにおいて正解以外の行にも高い関連度スコアを付与してしまう「過剰検知」が課題として明らかになった．さらに，モデルごとの特性差として，GPT-4o では数値の表層的な不一致を優先して関連なしと判断する傾向が，Qwen2.5 では特定のラベルにおいて過剰に抽出を行う傾向がそれぞれ確認された．

今後の課題としては，プロンプトの改良や関連度スコアの閾値の精査により，過剰検知を抑制し抽出の厳密性を向上させることが挙げられる．また，本実験で明らかになった手法間の特性を踏まえ，入力データの規模に応じて一括入力方式と 2 段階推論を動的に切り替えるハイブリッド型システムの構築も検討すべき重要な課題である．今後は，関連箇所が複数行に跨るケースや離れた数行に点在する複雑なデータセットへの拡張を行い，より高度な事実確認支援システムの実現を目指す．

## 謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

## 文　　献

[1] 黒川博生, 宮森恒. 大規模言語モデルを用いた文書補強とリランキングによる統計データ検索. 情報処理学会論文誌データベース (TOD), Vol. 18, No. 3, pp. 20–34, 2025.

[2] 宮﨑隆豪, 宮森恒. 統計データを用いた事実確認支援のための統計データ内の関連箇所抽出. In *DEIM2025*, pp. 8f–02, Kyoto, Japan, 2025. Information Processing Society of Japan.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 64–77, 2020.

[6] Dheeru Dua, Odin Wang, Aniruddha Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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)*, pp. 2368–2378, 2019.

[7] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, Online, July 2020. Association for Computational Linguistics.

[8] Jonathan Herzig, Paola Spangher, Jonathan Bogin, Ronen Chen, and Jonathan Berant. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 11270–11285. Association for Computational Linguistics, 2020.

[9] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6859–6866, 2019.

[11] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Chao Li, and Maosong Sun. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 892–901, 2019.

[12] Pritam Deka and Ashwathy Revi. PD-AR at ArAIEval shared task: A BERT-centric approach to tackle Arabic disinformation. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pp. 570–575, Singapore (Hybrid), December 2023. Association for Computational Linguistics.

[13] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. Overview of the NTCIR-15 data search task. In *Proceedings of the NTCIR-15 Conference*, 2020.

[14] OpenAI. Gpt-4o technical report. https://openai.com/research/gpt-4o, 2024. Accessed: 2024-08-05.

[15] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

# 事実確認支援のための言説と統計データ関連箇所との整合性検証

樫山　和貴†　　宮森　　恒†

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
E-mail: †g2253280@cc.kyoto-su.ac.jp

**あらまし**　本稿では，SNS 等における言説の真偽判定を支援するため, 統計データを活用して言説と統計データの関連箇所の整合性を検証する問題に取り組む. 従来の手法では，統計データの構造や暗黙的な意味関係を捉えることができず，キーワード検索や一般的なテキストベースの自然言語推論に依存しており正確な判断が難しいという課題があった．本研究は，統計データを活用し, 言説と統計データ間の意味的関係を捉えた整合性を検証する． 本稿では，言説テキストと統計データ関連箇所を用いた多クラス含意関係認識の問題として定式化する．言説テキスト, 関連箇所, 含意関係ラベル（含意/矛盾/部分的に真/判定不能）の 3 つ組からなるデータセットが不可欠であるが, 既存研究では提供されていない．そこで, 統計データと関連箇所から LLM を用いて SNS 投稿を模した言説テキストを生成し，人手による精査と修正を加えて含意関係ラベルを付与することで，数千件規模のデータセットを構築する．実験では, 構築したデータセットを用いて統計的推論に対応した含意関係認識モデルの性能を評価する．その結果，本手法が統計データに基づく事実確認において高い整合性検証能力を持つことを示す.

**キーワード**　事実確認支援，統計データ検索，整合性検証，含意関係認識，ソーシャルメディア

## 1　はじめに

現代社会において，SNS では，情報共有やコミュニケーションの主要な手段として広く利用されている．その利便性と迅速性から，個人や組織が情報を発信し，受け取るスピードは劇的に向上した．その一方で，誤情報やデマが拡散しやすいという問題が顕在化しており，統計データや事実関係を歪めた情報が拡散されることで，社会的な混乱や誤解が生じ，時には政策決定や個人の判断に重大な影響を与えることがある．近年，SNS 上では，新型ウイルスの感染拡大に関する誤情報が急速に拡散し，ワクチン接種率の低下や不安の増大を招いた事例がある．誤情報は信頼性の低い出所や意図的な操作による場合が多く．正しい情報との区別が難しいことが課題であり，こうした問題を解決するため，事実確認支援システムの開発が必要である．特に, 統計データを基にした客観的な検証は，情報の信頼性を高めるために重要である．以上の背景から，本研究では，SNS 等の投稿内容と公的な統計データとの整合性を検証するためのデータセットを構築し，構築したデータセットを用いて整合性検証を検証する.

統計データを用いた整合性検証は，事実確認の客観性を確保するための重要な手法である．統計データは，政府機関や研究機関などの公的機関によって収集・公開される信頼性の高い情報であり，一定の方法論に基づいて収集・分析されているため，検証基準として活用することで主観的な判断を排除することができる．また，経済指標や人口動態などの統計データは，信頼性が高いだけでなく，広範囲にわたるテーマを包括しているため，さまざまな分野での整合性検証に適用可能である．統計データを基盤とした検証は，個人ユーザーが情報の信憑性を判断する際の大きな助けとなる．SNS やニュースサイト上で多種多様な情報が飛び交う中，信頼性の高い統計データを基準として活用することで，ユーザーは主観的な判断に頼ることなく，情報の正確性を効率的に確認できる.

データセット構築の基盤として，NTCIR-15 が提供する統計データを活用する．NTCIR-15 は信頼性の高い統計データを提供しており，整合性検証のためのベースとなるデータとして適切である．SNS 上の投稿内容には，膨大かつ多様な情報が含まれており，それらの正確性を検証することは大きな課題である．本研究では，現実の SNS 投稿を直接使用するのではなく，NTCIR-15 の統計データを基に，SNS 上に見られるような書き込みテキストを生成するアプローチを採用した．この手法により，統計データとの直接的な整合性検証が可能となり，研究の効率性と信頼性を高めることができる.

本稿の目的は，SNS 等における言説に対し，統計データを用いてその真偽を判断する支援を行うことである．SNS は情報共有の重要な場である一方で，誤情報や誤解を招く投稿が拡散されるリスクも高い．これにより社会的混乱が生じることから，真偽を客観的かつ効率的に検証する仕組みが求められている．本研究では，公的に信頼のおける統計データを基盤とし，SNS 投稿の内容と統計データの間に整合性があるかを検証するモデルを構築することを目指す.

本提案手法では，信頼性の高い NTCIR-15 の統計データを基盤とし，大規模言語モデルを活用して，SNS 投稿を模倣したデータセットを生成する手法を採用する．具体的には，統計データを基に「統計データを支持する言説（True)」「統計データと矛盾する投稿（False)」「統計データを部分的に支持する言説 (Partly_true)」「統計データから判定不能な言説（Undeterminable)」の 4 つのカテゴリに分類される投稿データを構築する．この生成プロセスにより，統計データとの整合性を直接的

に検証できるデータを生成することが可能となる．

次に，生成された言説を評価し，その内容の適切性および信頼性を詳細に分析する．評価基準としては，情報の正確性，論理的一貫性，表現の明瞭性を設定し，特に統計データとの整合性を重視する．具体的には，言説が統計データに基づいて正確に記述されているか，ならびに誤解を招く表現や矛盾が含まれていないかを精査し，検証を行う．この評価を通じて，生成された言説の品質を把握し，統計データに基づく適切な情報提供の可能性を検討する. さらに，評価結果を分析し，言説の信頼性向上に向けた課題を明確にすることで，より精度の高い事実確認支援の実現を目指す.

本研究の特徴は，NTCIR-15 の統計データを基に，「言説テキスト」「統計データ関連箇所」「含意関係ラベル（True/False/partly_Ture/Undeterminable）」の 4 つからなるデータセットを構築し，4 つの多クラスラベルに対して，表形式データを維持したまま整合性検証を行うことである．このデータセットは，含意関係認識モデルの訓練および評価の基盤として機能し，SNS 投稿が統計データを支持する（True），矛盾している（False），部分的に支持する（Partly_True）, 判定不能（Undeterminable）を分類する能力を正確に測定することを可能にする．これにより，提案手法が SNS 投稿の真偽判定に有効であるかを明らかにし，SNS 上の情報信頼性向上に貢献することを目指している．

## 2 関連研究

### 2.1 自動化ファクトチェック

近年フェイクニュースの拡散防止を目的とした自動ファクトチェックの研究が活発化している. [11] が提唱した枠組みをはじめとして自動ファクトチェックのプロセスは，主に 4 つのタスクで構成されている．第一に「主張検出」であり，ドキュメント内から検証すべき重要な主張を抽出する． 第二に「証拠の取得」であり信頼性の高いデータベースや，統計情報，Wikipedia などから主張に関連する証拠を収集する．この段階では，従来のテキスト情報のみならず，画像や動画などのマルチモーダルデータを証拠として扱う研究も進展している． 第三に「判定予測」であり，収集した証拠と主張を突き合わせ含意関係認識などの技術を用いて真偽を判断する． 最後に「根拠生成」であり，なぜそのような判定に至ったかの理由を提示する．この一連のプロセスにより，説明可能で透明性の高いシステムの実現が目指されている．自動化ファクトチェックでは多くの研究が「支持」「反論」「情報不足」の 3 値分類あるいは単純な「真」「偽」の 2 値分類に焦点を当ててきた．しかし，現実では必ずしも白黒が明確ではなく，一部の情報が正しいものの誤解を招く表現や文脈に依存する複雑な構造を含んでおり，単純なラベル付では微細なニュアンスに対応できない．また，既存のデータセットには特定の単語が含まれると「偽」になりやすいといったバイアスが含まれておりこれがモデルの汎化性能に悪影響を与える可能性も示唆されている [5] こうした背景からよりきめ細やかな判定を可能にする新たなラベル付け手法やフレームワークが求められている． そこで本研究では，言説と統計データの整合性を検証する際に単純な二値分類ではなく，「含意」「矛盾」「部分的含意」「判定不能」の 4 クラスを用いたラベル付けを行う．これにより統計数値とテキスト間の複雑な意味関係を精緻に捉え，より高精度かつ実用的なファクトチェック判断の実現を目指す. .

### 2.2 表を用いたファクトチェック

従来のファクトチェック研究の多くは，Wikipedia の記事やニュース記事などの非構造化テキスト（自然言語文）を主要な証拠源として利用してきた．しかし，世界中の Web 上には，統計データ,，スポーツの試合結果など，豊富な情報が表形式やデータベースなどの構造化データとして存在している．これらの構造化データは従来のテキスト中心のアプローチでは検証が困難であった数値的な主張や比較を含む主張に対して不可欠な証拠を提供する．そのため表形式データを活用したファクトチェックは検証範囲の拡大と精度の向上を実現する新たなアプローチとして注目されている. [3] 表形式データを用いた検証は通常のテキスト処理よりも高度な推論能力を必要とする．テキストデータが主に言語的な意味理解を要するのに対し表データは行や列の構造を理解した上で記号的推論を組み合わせる必要があるためである．Wikipedia の表に基づいた真偽判定タスクである TabFact や，テキストと表の両方を証拠として扱う FEVEROUS の提案は，構造化データの理解における新たな挑戦領域を生み出している. [12] [1] [4] こうしたデータセットの登場に伴い検証モデルの研究も急速に進展し，SQL クエリの実行履歴を通じて表構造を学習する TAPEX を提案し，TabFact における SOTA 達成を通じて，表をフラットなテキストとして扱う従来手法に対する構造的理解の重要性を実証した. [7] 一方で，膨大なテキストデータによる事前学習を通じて高度な汎用推論能力を獲得した LLaMA や GPT シリーズに代表される大規模言語モデル（LLM）の発展も，この分野に大きな影響を与えている. [10] しかし既存手法の多くは依然として「真」か「偽」かの二値分類に焦点を当てており，統計データが持つ複雑な含意関係を十分に扱えていない場合がある．本研究では TAPEX のような構造理解に優れたモデルや Llama 等の LLM が持つ高度な言語生成・理解能力を活かしつつ，統計データに基づいて生成されたテキストとの整合性を多クラスラベルを用いてる検証する.

### 2.3 統計データを用いたデータセット構築

統計データなどの構造化データを自然言語記述に変換する Data-to-Text タスクは，情報の信頼性と説明性を向上させる重要な手段として注目されている．特に，大規模なデータセットから得られる有用な情報を，一般のユーザーが理解しやすい形で表現できる点は大きな利点である． しかし，表データの多様な形式への対応や，文脈に応じた適切な数値情報の選択，過不足のない要約，さらには生成文の流暢さと事実への忠実性の両立など，解決すべき課題は多岐にわたる．「WikiStatCells」では，Wikipedia 記事と統計データとの正確な対応関係をアノ

テーションすることにより，統計データを基にしたテキスト生成モデルの学習や評価，および統計データ検索タスクに貢献している.. [13] [2] また，「ToTTo」では選択されたセル情報に対して忠実な説明文を生成するタスクを提案しており，構造化データの活用に関する研究は深化している. [9] しかし，既存のデータセットの多くは Wikipedia のような解説的な長文テキストを主な対象としており，SNS 上の投稿に見られるような主観的，あるいは断定的な短文言説とは性質が異なる． そこで本研究では，NTCIR-15 の統計データを基盤として，SNS 投稿を模倣したデータセットの構築を行う． 具体的には，統計データとの整合性に基づき，単純な二値分類ではなく，「含意」「矛盾」に加え，「部分的真」「判定不能)」の計 4 クラスのラベルに対応する言説を生成する．これにより，より複雑で現実的なシナリオにおける整合性検証を目指す.

## 3 データセットの構築

本研究の目的である統計データに基づいた言説の整合性検証を行うためには，信頼性の高い統計データと，それらを参照して生成された多様な言説（含意・矛盾・部分的に真・判定不能)，および正解ラベルが紐付いた大規模なペアデータが必要不可欠である．しかし，既存のデータセットの多くは Wikipedia のような解説文を対象としており，SNS 上で見られるような主観的かつ断定的な短文言説と統計データを体系的に結びつけたものは存在しない．そこで本研究では，NTCIR-15 の統計データをソースとして，大規模言語モデルを用いた自動生成プロセスにより，SNS 投稿を模した言説データセットを構築する．本データセットの構築プロセスは，主に「統計データの前処理」,「関連箇所の特定」,「言説生成とラベリング」の 3 つのから構成される．以下に各手順の詳細を述べる.

### 3.1 統計データの前処理

本研究における言説生成では，大規模言語モデルを活用する．しかし，現行の LLM には一度に入力可能なトークン数に厳格な上限が存在する．そのため，NTCIR-15 で提供されるような M 行 N 列の巨大な統計データ全体を，そのまま LLM に入力して分析や言説生成を行うことは不可能である．このような技術的制約の下では，入力情報の質と量を最適化するための前処理が不可欠となる．単にデータを切り捨てるのではなく，限られた入力枠内で，いかにして元データの本質的な情報を LLM に伝達するかが極めて重要である．特に，行数 M が大きい場合，後段の処理における計算コストの増大や，情報過多による多重共線性といった問題を引き起こす可能性がある．そこで本提案手法では，これらの課題に対処し，データに内在する構造を抽出するために，前処理として次元削減を適用する．具体的には，元の M 行 N 列のデータ行列を，N 行 N 列程度のサイズに縮小・分割し，複数の扱いやすい統計データとして再構成する.

### 3.2 関連箇所の特定と文脈化

大規模言語モデルを用いた言説生成において，入力プロンプトに含まれる情報の質と密度は，生成されるテキストの正確性および論理的整合性を決定づける最も重要な要因である．広範かつ多岐にわたる統計データ全体を入力するのではなく，分析の核となる「関連箇所」を的確に特定し，LLM が焦点を絞って解釈できる形式で提示するプロセスが不可欠となる．本研究では，LLM に「どのデータを見て」「どのような文脈で」判断すべきかを明確に指示するため，以下の手順を用いて関連箇所を特定し，言説生成のためのコンテキストを構築する.

1. **データチャンクの選定**
   前節の前処理によって分割・縮小された複数のデータチャンク群（例：「統計データ名_列数_連番.csv」）の中から，分析対象とするファイルをランダムサンプリングによって選定する．ファイル名に含まれるメタデータは，そのデータが元データのどの部分に由来するかを示す情報として保持する.
2. **対象行の指定**
   次に，選定されたデータチャンク内に含まれるレコードの中から，言説生成の主たる対象となる特定の 1 行をランダムに指定する．統計表には多数の項目が含まれるが，LLM に対して「この行のデータに着目せよ」という明確なアテンションを与えることで，生成される言説の主題を固定する.
3. **周辺行による文脈の付与**
   指定されたターゲット行単体の数値情報のみでは，その値が全体の中で高いのか低いのか，あるいは時系列変化の中でどのような位置にあるのかといった「相対的な傾向」を LLM が理解することは困難である．実際，予備調査において周辺行を含まない設定（$k = 0$ または $k = 1$）で生成を試みたところ，順位関係や前後の文脈を無視した言説が生成される傾向が見られた．そこで本手法では，ターゲット行を中心として，その前後 5 行ずつ（計 11 行）を一つの意味的なまとまりとして切り出す処理を行う.
4. **メタデータの結合と構造化**
   切り出した行データに対し，表のヘッダー情報や単位といったメタデータを明示的に結合する．これにより，単なる数値の羅列を，人間が読むのと同等の「意味のある統計情報」へと変換する.

### 3.3 言 説 生 成

LLM を用いた言説生成においては，生成されたテキストに対し，「True」「false」「Partly_True」「undeterminable」の 4 種類の正解ラベルを付与する．ここで，各ラベルはそれぞれ「含意」「矛盾」「部分的含意」「判定不能」に対応するし，ラベリングは言説生成と同時に LLM に行わせる．プロンプトを通じて，統計データの内容と言説の整合性を生成時に検証させることで，各言説に対して論理的に適切なラベルが分類・付与されるように設計する.

## 4 提案手法

### 4.1 問題設定

本研究におけるタスクは，統計データ（表形式データ）の関連箇所 $T$ と，それを根拠としているとされる言説テキスト $S$ を入力とし，両者の意味的な整合性ラベル $y$ を予測する多クラス分類問題として定式化される．

具体的には，関連箇所 $T$ と言説テキスト $S$ のペアに対して，最適なクラス $y \in \mathcal{Y}$ を出力する予測モデル $f$ を構築することを目的とする．

$$y = f(T, S) \tag{1}$$

ここで，分類対象となるクラスラベルの集合 $\mathcal{Y}$ は，以下の 4 つのカテゴリにより定義される．

- **True (含意)** ($y_{true}$): 言説 $S$ の内容が統計データ の関連箇所 $T$ によって論理的に支持される．
- **False (矛盾)** ($y_{false}$): 言説 $S$ の内容が統計データの関連箇所 $T$ と矛盾する，あるいは事実と異なる記述を含む．
- **Partly_True (部分的真)** ($y_{part}$): 言説 $S$ の一部は統計データの関連箇所 $T$ によって支持されるが，一部は支持されない，あるいは不正確な記述を含む．
- **Undeterminable (判定不能)** ($y_{und}$): 統計データの関連箇所 $T$ の情報のみからは，言説 $S$ の真偽を論理的に導出できない．

検証対象となる関連箇所 $T$ には，NTCIR-15 に含まれる信頼性の高い公的統計データを用いる [6]．本来，この予測モデル $f$ の学習には，実際の Web 上の言説と統計データが紐付いた大規模なデータセット $\mathcal{D} = \{(T_i, S_i, y_i)\}_{i=1}^{N}$ が必要となる．しかし，現状では統計データを引用した Web 上の言説は体系的に整理されておらず，教師データとして利用可能なリソースが著しく不足している．

そこで本研究では，NTCIR-15 の統計データを基に大規模言語モデルを用いて生成した「疑似的な SNS 投稿」を検証対象とする．すなわち，本研究の問題設定は，生成された言説 $S$ が元の統計データの関連箇所 $T$ に対して正しい含意関係 $y$ を持っているかを，提案モデルがいかに正確に識別できるかを評価することにある．

### 4.2 整合性検証

#### 4.2.1 TAPEX の採用

統計データのような構造化データと自然言語テキストの間の整合性を検証するためには，表の構造を正しく理解し，数値の比較や集計といった推論を行う能力が不可欠である．一般的な大規模言語モデルは高い言語能力を持つ一方で，複雑な表構造の認識や厳密な数値計算を苦手とする場合がある．そこで本研究では，整合性検証モデルの基盤として，表形式データの理解に特化した事前学習済みモデルである TAPEX を採用する．TAPEX は，表に対する SQL クエリの実行結果を予測するというタスクで事前学習されており，これにより表の構造や数値間の関係性を深く理解する能力を獲得している．この特性は，統計データに基づいた事実確認タスクにおいて極めて有効であると考えられる．

#### 4.2.2 学習データの入力形式

TAPEX は Transformer ベースのエンコーダ・デコーダモデルであり，入力として 1 次元のテキストシーケンスを要求する．そのため，2 次元構造を持つ統計データを構造情報を欠損させることなく自然言語のシーケンスへと変換する処理が必要となる．本研究では，各セルを「ヘッダー名 : セルの値」のペアとして記述し，それらを特殊トークンで連結する．

さらに，モデルの推論精度を向上させるため，検証対象の言説が直接言及しているターゲット行の全セルに対して特定の記号を付与し，表内の重要なコンテキストを明示的に強調した．第二に，ターゲット行のみならず，その前後各 5 行を周辺コンテキストとして含めることで，最大 11 行の局所的な表構造を保持した状態で入力を行うこととした．

#### 4.2.3 ファインチューニング

構築した学習データセットを用いて，TAPEX に対して 4 クラス分類タスクのファインチューニングを行う．モデルは，入力された統計データと言説のペアから，両者の意味的関係が「True」「False」「Partly_True」「Undeterminable」のいずれであるかを予測するように学習する．汎用的な LLM にゼロショットで推論させるのではなく，統計データに基づく含意関係認識という特定のタスクに特化した学習を行うことで，特に「Partly_True」や「Undeterminable」といった複雑なニュアンスを含む関係性に対しても，高精度な判定能力を獲得することを目指す．以上のデータセット構築から TAPEX による整合性検証に至る，本提案手法の全体的な処理概要を図 1 に示す．

## 5 評価実験

### 5.1 実験目的

本実験の目的は，構築したデータセットおよび TAPEX を用いた提案手法が，統計データに基づく事実確認タスクにおいてどの程度有効であるかを定量的に検証することである．従来の一般的な言語モデルを用いた手法では，2 次元的な構造を持つ表データを 1 次元のテキスト列に変換して処理するため，行や列の対応関係や，セル間の階層構造といった重要な情報が失われる傾向にあった．これに対し，本研究で採用する TAPEX は，表構造を理解するための事前学習が行われている．本実験では，このモデルが統計データの構造的特徴を適切に認識し，数値の比較や集計といった高度な推論を必要とする整合性検証において，高い精度を発揮できるかを確認する．既存の多くのファクトチェック研究は「真」か「偽」かの二値分類に焦点を当ててきた．しかし，現実の言説には，一部は正しいが一部は誤っている「部分的真」や，提示されたデータだけでは判断できない「判定不能」といった複雑なケースが多々存在する．本実験では，提案手法がこれらの曖昧性を含む 4 つのラベル（True, False, Partly_True, Undeterminable）をどの程度正確に識別できるかを検証し，より精緻で実用的な事実確認支援が可能であることを示す．

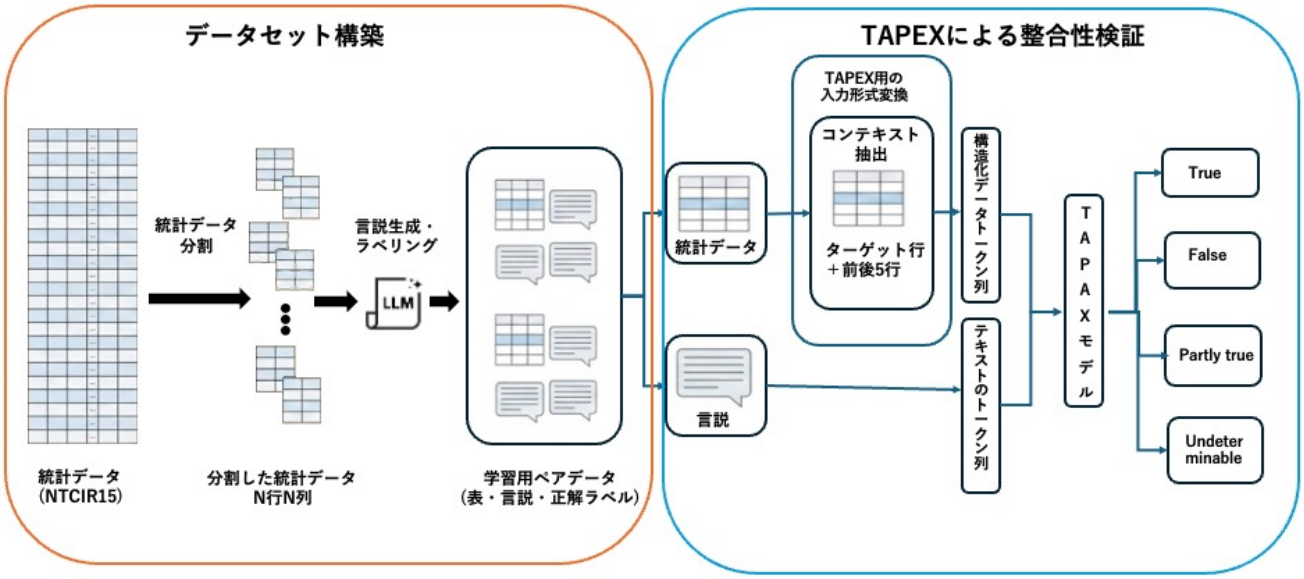



図 1　データセット構築および整合性検証の処理フロー

### 5.2 実験方法

本実験では，NTCIR-15 のデータセットから収集したデータに対し，第 3 章で述べた前処理を施した 500 件の統計データチャンクを使用した．これは，巨大な元データをそのまま使用するのではなく，モデルの入力制限および検証の焦点化を考慮し，意味的なまとまりを持つ単位（$N' \times N'$ サイズ）に再構成したものである．

これらの統計データチャンクをソースとして，GPT-4o [8] を用いて言説生成を行った．生成にあたっては，各統計データが示す数値情報に加え，関連箇所およびメタデータをプロンプトとして入力し，統計データとの整合性が 4 つのラベル（True, False, Partly_True, Undeterminable）のいずれかに該当する言説を出力させた．最終的に，合計 12,000 件の言説テキストを生成し，これを本実験のデータセットとした

### 5.3 比較手法

本研究の提案手法である TAPEX の有効性を検証するため，比較手法として汎用的な大規模言語モデルである Llama を採用する． 両手法とも入力データをトークン列として処理する点は共通しているが，表構造の取り扱い方に決定的な差異がある．

TAPEX は，入力トークンに対して行および列のインデックス情報を埋め込み表現として明示的に付与することで，表の構造情報を保持したままエンコードが可能である．

対照的に，Llama などの一般的な LLM は，表データを構造化データとして扱うための専用の入力層を持たない．そのため，比較実験における Llama への入力では，統計データを Markdown 形式や「ヘッダー名: 値」の形式等のテキストシーケンスに変換して入力する手法をとる．この場合，モデルはテキスト上の区切り文字や改行などのパターンから表構造を暗黙的に推論する必要がある．

本実験では，表の各セルの座標（行・列）を直接モデルに入力する TAPEX のアプローチと，テキスト形式に変換して汎用モデルの推論能力に委ねる従来的な LLM アプローチとの性能差を明らかにする．

### 5.4 データの分割と評価手順

構築した合計 12,000 件のデータセットを，学習用データとして全体の 80%にあたる 9,600 件，検証用データとして 10%にあたる 1,200 件，そして評価用データとして残りの 10%にあたる 1,200 件に分割した．

この際，学習データと評価データの間で，元となる統計データチャンクが重複しないように配慮し，統計データチャンクから生成された言説群は，すべて学習用か評価用のどちらか一方にのみ含まれるように分割を行った．これにより，モデルが統計データの内容そのものを丸暗記するのではなく，未知の統計データに対しても構造を理解し，汎化性能を発揮できるかを厳密に検証できる設定とした．

さらに，データの偏りによる影響を排除し，実験結果の信頼性と堅牢性を担保するため，上記の「データのランダム分割」および「学習・評価」の一連のプロセスを 4 回繰り返した．最終的な評価には，これら 4 回の試行における評価指標の平均値を採用する．

### 5.5 学習と評価指標

学習フェーズでは，前節で分割した学習用データ 9,600 件を用いて各モデルをファインチューニングし，4 クラス分類の推

論能力を学習させた．また，検証用データ 1200 件は，学習過程におけるモデルの挙動監視およびハイパーパラメータの調整に用いた． データのランダム分割によるバイアスを排除するため，この学習プロセスは分割手順と同様に 5 回独立して実施した．

続く評価フェーズでは，学習および検証には一切使用していない 600 件の評価用データを各試行のモデルに入力し，推論を行わせた． モデルが出力した予測ラベルと，データセット生成時に付与された正解ラベルとを比較し，各ラベルごとの適合率（Precision），再現率（Recall），F1 値を算出した． 最終的な性能評価には，これら 4 回の試行で得られた各指標の平均値を用い，提案手法の有効性を定量的に評価した．

### 5.6 評価結果

#### 5.6.1 混同行列を用いた誤分類の傾向分析

モデルの誤分類における傾向を詳細に分析するため，表 1 および表 2 に混同行列を示す．本行列は，行に正解ラベル，列に予測ラベルを配置し，4 回の試行における各ラベルの平均件数を算出したものである．

分析の結果，TAPEX は Llama と比較して，テーブル構造に基づいた極めて精緻な推論を行っていることが明らかになった．

正解が「True」の事例において, TAPEX が「Undeterminable」と誤認した件数はわずか 0.5 件にとどまった．これは，対照的な結果となった Llama（7.0 件）の 14 分の 1 という極めて低い数値である．Llama は検証対象が表内の多岐にわたる場合に情報の集約に失敗し，Undeterminable を選択する傾向があったのに対し，TAPEX は広範囲のセルを参照する必要がある場合でも，必要な根拠を的確に抽出できていた．特に，2 箇所以上のセルを横断的に参照する推論において，Llama は推論の複雑化に伴い確信度を低下させ「Partly_True」や「Undeterminable」へと分類する傾向が見られたが，TAPEX はテーブル構造に特化した事前学習の恩恵により，安定した判定を維持していた．

一方で，両モデルに共通して見られた課題として，数値の桁数が大きい場合に正解ラベルを正しく識別できず，他のラベルへ誤分類してしまう傾向が挙げられる．特に，正解が「False」であるにもかかわらず他のラベルへと判断してしまう事例において，対象となる数値が 6 桁以上の大きな値である場合，モデルが一部の桁の不一致を無視し，全体的な整合性のみで判定を下してしまう傾向が確認された．

TAPEX においてはこの傾向が顕著であり，正解が「False」の事例に対して True（27.8 件）といった他のラベルへ誤分類するケースが一定数発生している．これは，対象文中の数値が表内の値とわずかに異なる場合や，列数の多い大規模な表において一部の数値のみが改変されている場合に発生しやすい．

以上の分析から，TAPEX は表のレイアウト把握や論理構造の特定には極めて長けているものの，大きな桁数を含む微細な数値的不一致を厳密に識別できず，他のラベルを割り当ててしまう性質があることが示唆された．今後は，表構造の理解力を維持しつつ，数値の厳密な比較能力をいかに向上させるかが課題である．

表 1　TAPEX の推論性能に関する混同行列

| 正解ラベル | TAPEX による予測ラベル | | | |
|---|---|---|---|---|
| | True | False | Partly_True | Undeterminable |
| True | 279.0 | 14.2 | 6.2 | 0.5 |
| False | 27.8 | 271.0 | 0.8 | 0.5 |
| Partly_True | 3.0 | 1.2 | 294.0 | 1.5 |
| Undeterminable | 1.2 | 0.2 | 0.8 | 297.8 |

表 2　Llama の推論性能に関する混同行列

| 正解ラベル | Llama による予測ラベル | | | |
|---|---|---|---|---|
| | True | False | Partly_True | Undeterminable |
| True | 268.8 | 15.8 | 8.5 | 7.0 |
| False | 22.8 | 269.8 | 2.5 | 5.0 |
| PPartly_True | 10.0 | 4.0 | 278.0 | 7.8 |
| Undeterminable | 4.2 | 6.0 | 3.8 | 286.0 |

#### 5.6.2 定量評価によるモデル比較

表 3 に，TAPEX と Llama を用いた場合のクラス別の適合率，再現率，F1 値を示す． 実験の結果，全てのクラスにおいて TAPEX が Llama を上回る F1 値を達成した．特に「Undeterminable」および「Partly_True」クラスにおいては，TAPEX は 0.97 を超える極めて高い F1 値を記録しており，表構造を理解する能力の高さが示唆された．

一方で，Llama（表 3）は全てのクラスで F1 値が 0.88〜0.95 の範囲に留まった．特に「True」クラスの F1 値は 0.887 と最も低く，事実性が明確なデータであっても，TAPEX と比較して正答率が劣る傾向が見られた． 全体を通して，表形式データに基づく含意関係認識タスクにおいては，事前学習段階で表構造を学習している TAPEX の方が，汎用 LLM である Llama よりも高い適性を持つことが定量的に示された．

表 3　TAPEX と Llama の性能比較

| ラベル | 適合率 | | 再現率 | | F1 値 | |
|---|---|---|---|---|---|---|
| | TAPEX | Llama | TAPEX | Llama | TAPEX | Llama |
| True | 0.897 | 0.879 | 0.930 | 0.895 | 0.913 | 0.887 |
| False | 0.945 | 0.912 | 0.903 | 0.899 | 0.923 | 0.906 |
| Partly_True | 0.974 | 0.949 | 0.980 | 0.927 | 0.977 | 0.938 |
| Undeterminable | 0.991 | 0.935 | 0.992 | 0.953 | 0.992 | 0.944 |

### 5.7 考　　察

定量評価（表 3）において，TAPEX が全指標で Llama を上回った主要因は，事前学習における「表形式データとクエリの整合性」の学習量および、表構造を保持したエンコード手法にあると考えられる．特に「Partly_True」および「Undeterminable」における F1 値が 0.97 を超えている点は注目に値する．Llama のような汎用言語モデルは，表を線形なテキストとして処理するため，行・列の対応関係が複雑な場合に情報の欠落が生じやすい．これに対し，TAPEX は表の二次元的な構造を保持したまま処理を行うため，複数のセルを跨ぐ条件参照においても論理的な一貫性を維持できたと推察される．これは，Llama が「True」や「false」の事例を他のラベルへ誤分類し，再現率を低下させている結果（表 2）とも整合する．

混同行列の比較から，Llama 特有の振る舞いとして「Undeterminable への回避」が顕著に見られた．正解が「True」の事例を「Undeterminable」と誤認した件数は，TAPEX が 0.5 件であるのに対し，Llama は 7.0 件に上る．これは，Llama が長大なコンテキストや複雑な数値関係を処理する際，根拠の特定に至らずに判断を放棄する傾向があることを示唆している．

対照的に，TAPEX は判断を回避せず積極的にラベルを割り当てる傾向があるが，一方で数値的な脆弱性も確認された．具体的には，正解が「False」である事例を「True」ラベルへ誤分類するケースが合計 27.8 件発生している．これらの誤分類の多くは，同一行内に存在する複数の数値のうち一部のみが書き換えられた事例や，6 桁以上の大きな桁数を持つ数値において微細な不一致を含む事例で発生している．

これは，TAPEX が表の行と列の対応関係，すなわち「どの行にどの列が紐付いているか」という広域的な構造把握には極めて長けている一方で，同一行内の特定のセルに対する「局所的な数値検証能力」に課題があることを示している．言い換えれば，TAPEX は対象文に関連する「該当行」を正しく特定できたことに引きずられ，その行内の詳細な数値的不一致を見落とし，安易に「含意（True）」等のラベルを選択している可能性が高い．

以上のことから，TAPEX は表構造に特化した事前学習により，汎用 LLM を大きく上回る推論の安定性を獲得しているものの，大きな桁数を含む数値変化を厳密に識別する能力については依然として改善の余地があると言える．

## 6 ま と め

本研究では，SNS 等における言説の真偽判定を支援することを目的とし，統計データを用いた言説の整合性検証手法を提案した．信頼性の高い公的統計データである NTCIR-15 を基盤とし，大規模言語モデルを活用して SNS 投稿を模した言説データセットを構築した．本データセットは，従来の二値分類では扱いきれなかった曖昧な言説に対応するため，「含意」「矛盾」「部分的真」「判定不能」の 4 クラスで定義されている．

実験では，表構造の理解に特化した事前学習済みモデルである TAPEX と，汎用 LLM である Llama を用いて比較評価を行った．定量評価の結果，TAPEX は全てのクラスにおいて Llama を上回る F1 値を記録し，特に「判定不能」や「部分的真」といった複雑な論理関係の識別において高い優位性を示した．これは，二次元的な表構造を保持したままエンコードを行う TAPEX の構造的理解力が，統計データに基づく事実確認において極めて有効であることを示唆している．

また，誤分類の傾向分析を通じて，モデルごとの推論特性を明らかにした．Llama が複雑な情報に対して判定不能と出力する傾向があるのに対し，TAPEX は安定した推論を行う一方で，同一行内の微細な数値の不一致を見落とすという特有の課題も浮き彫りとなった．今後の課題として数値の厳密な比較能力の向上である．考察で述べた通り，現在のモデルは表構造の特定には長けているものの，数値の微細な差異を許容してしまう性質がある．数値の完全一致を判定するための外部モジュールとの連携や，数値の改変に敏感な学習手法の導入を検討する必要がある．

## 謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

## 文　　献

[1] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pp. 1–13, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Oana Balalau, Simon Ebel, Théo Galizzi, I. Manolescu, Quentin Massonnat, Antoine Deiana, Emilie Gautreau, Antoine Krempf, Thomas Pontillon, Gérald Roux, and Joanna Yakin. Fact-checking multidimensional statistic claims in french. pp. 20–29, 2022.

[3] Tien-Duc Cao, I. Manolescu, and Xavier Tannier. Searching for truth in a database of statistics. *Proceedings of the 21st International Workshop on the Web and Databases*, 2018.

[4] 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*. Association for Computational Linguistics, 2020.

[5] Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. Factoring fact-checks: Structured information extraction from fact-checking articles. *Proceedings of The Web Conference 2020*, 2020.

[6] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. Overview of the NTCIR-15 data search task. In *Proceedings of the NTCIR-15 Conference*, 2020.

[7] 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, 2022.

[8] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and et.al AJ Ostrow. Gpt-4o system card, 2024.

[9] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, Online, November 2020. Association for Computational Linguistics.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[11] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, *Proceedings of the ACL 2014 Work-*

*shop on Language Technologies and Computational Social Science*, pp. 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[12] Jianshu Chen Yunkai Zhang Hong Wang Shiyang Li Xiyou Zhou Wenhu Chen, Hongmin Wang and William Yang Wang. Tabfact : A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, April 2020.

[13] 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, Vol. 1, No. 1, 3 2023.

# 統計データ検索における視覚的文書検索の有効性分析

福岡　啓人†　　宮森　　恒†

† 京都産業大学先端情報学研究科　〒 603-8555 京都府京都市北区上賀茂本山
E-mail: †{i2486200,miya}@cc.kyoto-su.ac.jp

**あらまし**　政府機関等が公開する統計データは，フェイクニュース対策としての事実確認（ファクトチェック）において重要な情報源である．しかし，統計文書は図表，注記，本文など複数要素の配置（レイアウト）に強く依存する．そのため，OCR，表構造解析，言語モデルを組み合わせた従来のテキストベース手法では，処理パイプラインの複雑化と誤りの段階的な蓄積が課題となる．本稿では，文書画像を直接埋め込む視覚的文書検索（Visual Document Retrieval）モデルである ColPali を統計データ検索タスクに適用し，OCR を用いない視覚的アプローチの有効性と限界を検証する．NTCIR-16 データセットを用いた評価実験では，BM25 によるキーワード検索や LLM を利用した既存手法と ColPali の検索精度（nDCG）を比較する．さらに，検索結果の分析を通じて，ColPali と比較手法の性能差を生む要因を検討する．以上により，Visual Document Retrieval が統計データ検索において，実装容易な証拠候補探索手段としてどの程度有用であるかを明らかにする．

**キーワード**　統計データ検索，Visual Document Retrieval，VLM，情報検索

## 1　はじめに

近年，ソーシャルメディアを介した情報発信の容易化に伴い，虚偽や誤解を招く情報が拡散し，個人および社会へ影響を及ぼす事例が増加している．この状況に対処するため，言説の妥当性を検証する事実確認（以降，ファクトチェック）の重要性が高まっている．ファクトチェックでは，主張と客観的データとの整合性を確認することが，真偽判断の重要な根拠となる．本稿では，客観的データとして政府機関等が公開する統計データの利用を想定し，統計文書のアドホック検索（以降，統計データ検索）に焦点を当てる．

本稿では，先行研究および NTCIR–16 Data Search 2 タスクの定義に倣い，統計文書（＝メタデータ＋統計データ本体）を，タイトルや概要などのメタデータと，図表・注記・本文等を含む統計データ本体（PDF や CSV 等）から構成される単位として扱う．メタデータは比較的短い一方，統計データ本体は PDF や CSV など多様な形式で提供される．とりわけ PDF では，見出し，注記，図表の配置といったレイアウトが意味理解に直結する．したがって，メタデータやタイトル中の語に依存した単純なキーワード検索では，根拠となる文書を的確に取得できない場合がある．

統計データ検索の精度向上に向けて，メタデータに加えて表ヘッダ情報を抽出して検索に用いる手法 [14] や，表データから補強文書を作成した上で言語モデルによりリランキングする手法 [15], [16] など，様々な研究が行われてきた．しかし，これらの多くは言語情報（メタデータやヘッダ）や数値の並びに焦点を当てており，PDF 内の図表の配置，注記と表の対応関係といった「見た目」に依存する手掛かりを十分に活用できていない．また，表のみに着目した研究が多く，図表や周辺要素を含めた扱いは限定的である．さらに，高精度を狙う既存手法では，OCR，表構造解析，LLM による補強文書，リランキングなどを組み合わせるパイプラインが採用されている．この場合，構築コストが高いだけでなく，各段で生じた誤りが後段へ伝播し，性能低下を招く可能性がある．

これらの課題に対し，文書を OCR でテキスト化せず，文書画像を直接埋め込んで検索する Visual Document Retrieval(以降，VDR) が注目されている．特に ColPali [3] は，文書の見た目（レイアウト）を保持したまま埋め込み表現を得て検索する枠組みであり，OCR なしに検索を行える．統計文書では，図表や注記，本文といった要素も解釈に影響する．したがって ColPali 系（本研究では ColQwen）のような視覚的アプローチは，表構造に限らない文書全体の視覚的特徴を手掛かりとしつつ，複雑な前処理を回避できる可能性がある．本研究では，この可能性を統計データ検索タスクで検証する．

本研究の目的は，統計データ検索というドメインに対して，ColQwen に基づく視覚的アプローチによる，有効性と限界を検証することである．本稿の貢献は以下のとおりである．

1. 統計データ検索に VDR を適用し，メタデータ検索において BM25 を上回る精度を示した点，
2. VDR による PDF 検索の限界（ページ単位エンコードによる文脈喪失）を明らかにした点，
3. 文脈喪失を補うためメタデータを PDF ページと同時に VLM へ入力する手法を試み，その効果が限定的である原因を分析した点

以下，本稿の構成を述べる．まず 2 節で統計データ検索および視覚的文書検索に関する関連研究を整理する．続いて 3 節で ColPali に基づく視覚的文書検索の枠組みと，その実装として ColQwen を用いた本研究のインデキシングおよびスコアリングを示す．次に 4 節でデータセットと比較手法を述べ，ベースラインと提案手法の実験の結果を報告する．最後に 5 節で考察

を行い，6 節でまとめと今後の課題を述べる．

## 2 関連研究

### 2.1 統計データと表理解

統計文書は，図表，注記，および本文が密接に関連し合う複合的な文書である．したがって，統計データ検索においては，表内のセル構造だけでなく，ページ内における各要素の配置や関係性といった視覚的文脈全体を捉えることが不可欠である．

これまで，表理解の分野では，TAPAS [4] や TaBERT [13] のように，表の構造情報（行・列）を埋め込みとして学習するモデルが主流であった．また，統計データ検索に特化した研究においても，表ヘッダを抽出してメタデータを補強する手法 [14] や，表データから説明文を生成して LLM で活用する手法 [15], [16] など，「いかに表構造を正確にテキスト化するか」に焦点が当てられてきた．

しかし，これらのアプローチは「統計文書＝表」という前提に立つものが多く，表構造以外の視覚要素（グラフの傾向，注記と表の空間的な対応など）を十分に活用できていない．また，複雑なレイアウトを持つ PDF から表領域のみを正確に切り出し，構造解析すること自体が困難であり，その前処理の誤りが検索精度を低下させる要因となる．本研究では，統計文書を表構造だけに限定せず，視覚的な文書全体として埋め込むことで，図表や注記を含む包括的な情報を用いた検索の有効性を検証する．

### 2.2 視覚的文書理解と検索

近年，レイアウトや図表を含む文書画像を対象とした視覚的文書理解の研究が急速に進展している．既存手法は，テキスト情報の扱いに基づき，(i) OCR 併用型と (ii) OCR レス型に大別される．OCR 併用型の代表例である LayoutLM シリーズ [5], [11], [12] は，OCR を用いて抽出したテキストとレイアウト情報を統合する手法である．文字情報を活用できる一方，性能は前段の OCR 品質に強く依存する．これに対し，Donut [8] や Pix2Struct [10] といった生成モデルによって確立された OCR レス型のアプローチは，文書全体を画像としてエンコードすることで，誤認識に起因するノイズを抑制し，視覚情報を保持したままエンドツーエンドでの処理を可能にした．

本研究では，この OCR レス型のパラダイムを文書検索（Visual Document Retrieval）へと応用した ColPali [3] に着目する．ColPali は，ColBERT [7] の Late Interaction 機構を VLM（Vision Language Model）へと拡張したモデルである．生成モデル（Donut 等）がテキスト生成を目的とするのに対し，ColPali は検索ランク付けのためのスコアリングに特化している．具体的には，クエリをテキストトークンの列，文書を画像パッチの列として，それぞれ多次元の埋め込み表現（マルチベクトル）に変換する．

先行研究では，ColPali 系の VDR を統計データ検索に適用した検討は報告されていない．そこで本研究では，統計データ検索タスクに ColQwen（ColPali の派生モデル）を適用し，OCR レスの VDR アプローチの有効性と限界を評価する．

## 3 提案手法

### 3.1 利用するモデル

本研究では，視覚的文書検索を実現するモデルとして ColPali およびその派生モデルである ColQwen に着目する．ColPali [3] は，視覚言語モデル（VLM）である PaliGemma-3B [2] をバックボーンとして採用した VDR モデルである．一方，ColQwen2.5 は，ColPali のアーキテクチャを踏襲しつつ，ベースモデルを Qwen2.5-VL [1] に変更した派生モデルである．PaliGemma は多言語データを含む大規模コーパスで事前学習されている一方で，日本語対応が明示されていない．これに対し，Qwen2.5-VL は多言語データセットで学習されており，日本語を正式にサポートしている．本研究で扱うデータセットは英語であるものの，将来的な日本語統計文書への適用や言語非依存な検索モデルの構築を見据え，本実験では多言語性能に強みを持つ ColQwen を採用する．

### 3.2 問題設定

クエリ集合を $Q = \{q_i\}$，統計文書集合を $D = \{d_j\}$ とする．クエリ $q \in Q$ は，1 回の検索で与えられる語列 $q = (w_1, \ldots, w_{|q|})$ で表す．各統計文書 $d_j \in D$ は，検索に利用可能なメタデータ $m_j$ と，1 つ以上の統計データ本体 $\{c_{j,k}\}_{k=1}^{n_j}$ の組として表す．メタデータ $m_j$ は，タイトル $t_j$ と説明文 $desc_j$ から構成される．

$$d_j = (m_j, \{c_{j,k}\}_{k=1}^{n_j}), \quad m_j = (t_j, desc_j) \tag{1}$$

ここで $n_j$ は文書 $d_j$ に含まれる統計データファイルの数であり，各 $c_{j,k}$ は PDF または CSV ファイルに対応する．PDF ファイル $c_{j,k}$ が複数ページから構成される場合，各ページを $p_{j,k,l}$ $(l = 1, \ldots, P_{j,k})$ で表す．ここで $P_{j,k}$ はファイル $c_{j,k}$ のページ数である．あるクエリ $q$ と文書 $d$ に対してスコア $s_{q,d}$ を割り当てるランキング関数を $f(q, d)$ とする．

$$s_{q,d} = f(q, d) \tag{2}$$

統計文書のアドホック検索は，クエリ $q$ に対して $D$ 中の文書をスコア降順で並べたランキングリスト $\pi_q$ が適切となるように $f$ を定める問題である．

$$\pi_q = \mathrm{sort}_{d \in D}(D;\, s_{q,d}) \tag{3}$$

### 3.3 提案手法の定式化

本研究では，ColQwen [3] に基づく VLM を用いて，統計文書の視覚的表現とテキスト表現を統合した検索を行う．ColQwen は，テキストエンコーダ $f_{\text{text}}$ と画像エンコーダ $f_{\text{vision}}$ から構成される．

統計文書 $d_j = (m_j, \{c_{j,k}\}_{k=1}^{n_j})$ において，メタデータ $m_j = (t_j, desc_j)$ のタイトル $t_j$ と説明文 $desc_j$，および統計データ $c_{j,k}$ をそれぞれエンコードする．タイトル，説明文，および CSV 形式の統計データはテキストエンコーダ $f_{\text{text}}$ を用いて埋め込み表現へ変換する．

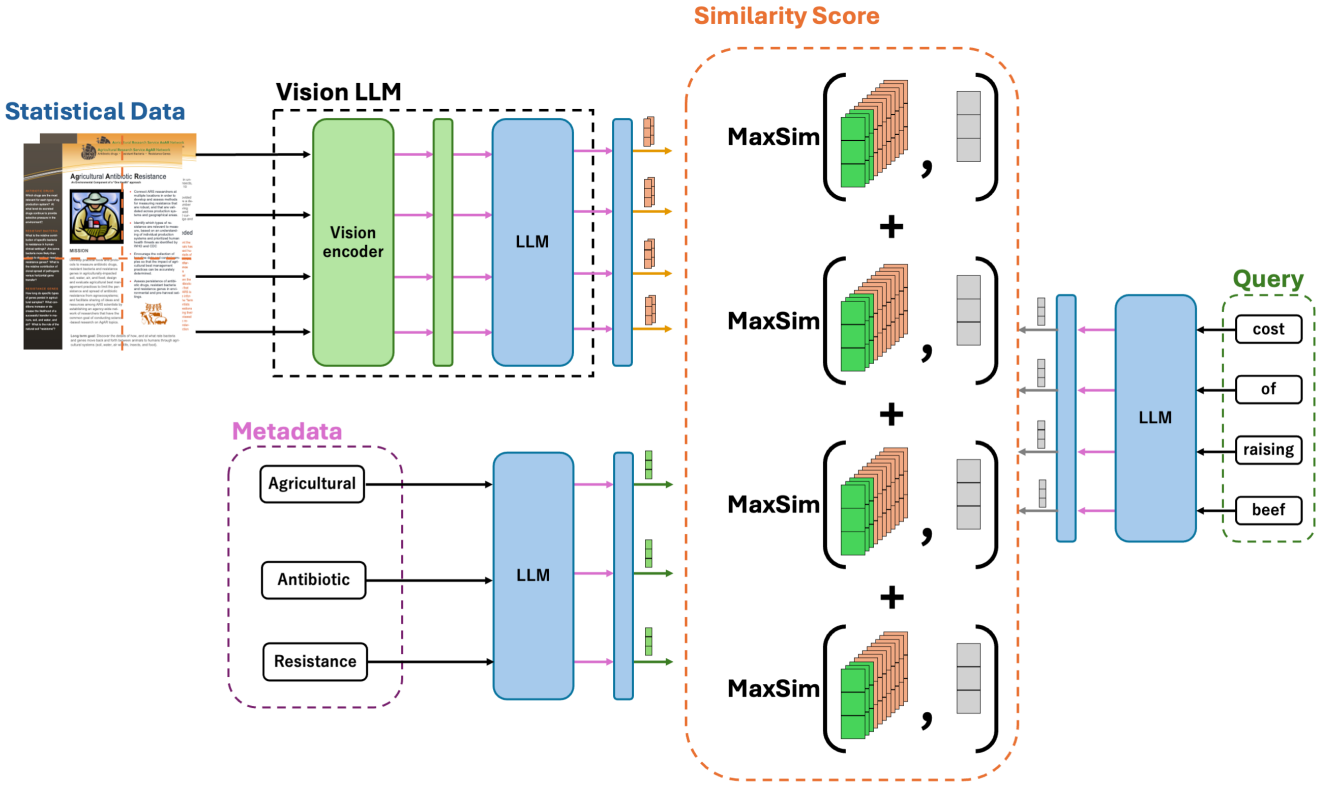



図 1　提案手法の概要．統計データが PDF の場合は Vision Encoder により画像パッチ列へ変換される．一方，CSV の場合はメタデータと同様に Vision Encoder を通さず，テキストとして直接 LLM へ入力される． 最終的にクエリとの MaxSim により類似度を算出する．

$$\mathbf{E}_{t_j} = f_{\text{text}}(t_j) \in \mathbb{R}^{N_t \times D} \tag{4}$$

$$\mathbf{E}_{desc_j} = f_{\text{text}}(desc_j) \in \mathbb{R}^{N_{desc} \times D} \tag{5}$$

$$\mathbf{E}^{\text{csv}}_{c_{j,k}} = f_{\text{text}}(c_{j,k}) \in \mathbb{R}^{N_{\text{csv}} \times D} \tag{6}$$

ここで，$N_t$，$N_{desc}$，および $N_{\text{csv}}$ はそれぞれタイトル，説明文，および CSV をトークン化した際のトークン数，$D$ は埋め込み次元数を表す．

一方，PDF 形式の統計データはページ画像として扱い，画像エンコーダ $f_{\text{vision}}$ を用いて画像パッチ単位の埋め込み表現へ変換する．PDF ファイル $c_{j,k}$ が $P_{j,k}$ ページから構成される場合，各ページ $p_{j,k,l}$ を個別にエンコードし，それらを連結する．

$$\mathbf{E}_{p_{j,k,l}} = f_{\text{vision}}(p_{j,k,l}) \in \mathbb{R}^{N_{j,k,l} \times D} \tag{7}$$

$$\mathbf{E}^{\text{pdf}}_{c_{j,k}} = \left[\mathbf{E}_{p_{j,k,1}}; \ldots; \mathbf{E}_{p_{j,k,P_{j,k}}}\right] \in \mathbb{R}^{\left(\sum_{l=1}^{P_{j,k}} N_{j,k,l}\right) \times D} \tag{8}$$

ここで，$N_{j,k,l}$ はページ $p_{j,k,l}$ の画像パッチ数を表す．

文書 $d_j$ 全体のマルチベクトル表現 $\mathbf{E}_{d_j}$ は，タイトル，説明文，および各統計データファイルの埋め込み表現を連結して得られる．

$$\mathbf{E}_{d_j} = \left[\mathbf{E}_{t_j}; \mathbf{E}_{desc_j}; \mathbf{E}_{c_{j,1}}; \ldots; \mathbf{E}_{c_{j,n_j}}\right] \in \mathbb{R}^{N_{d_j} \times D} \tag{9}$$

ここで，$N_{d_j}$ は文書 $d_j$ の総ベクトル数であり，$N_{d_j} = N_t + N_{desc} + \sum_{k=1}^{n_j} N_{c_{j,k}}$ で表される．$N_{c_{j,k}}$ は統計データファイル $c_{j,k}$ のベクトル数であり，PDF の場合は $\sum_{l=1}^{P_{j,k}} N_{j,k,l}$，CSV の場合は $N_{\text{csv}}$ である．

クエリ $q = (w_1, \ldots, w_{|q|})$ についても，テキストエンコーダ $f_{\text{text}}$ を用いてマルチベクトル表現へ変換する．

$$\mathbf{E}_q = f_{\text{text}}(q) \in \mathbb{R}^{N_q \times D} \tag{10}$$

ここで，$N_q$ はクエリのトークン数を表す．

クエリ $q$ と文書 $d_j$ 間のスコア $s(q, d_j)$ は，Late Interaction (MaxSim) [7] により算出する．具体的には，クエリの各トークン埋め込み $\mathbf{E}_q^{(i)}$ について，文書の全ての局所埋め込み $\mathbf{E}_{d_j}^{(j)}$ との内積を計算し，その最大値を求める．最終的なスコアは，全てのクエリトークンに対するこれらの最大値の総和として定義される．

$$s(q, d_j) = \sum_{i=1}^{N_q} \max_{j \in [1, N_{d_j}]} \left\langle \mathbf{E}_q^{(i)}, \mathbf{E}_{d_j}^{(j)} \right\rangle \tag{11}$$

このスコアリング方式により，クエリの各概念が文書内のどの局所領域（テキストトークンまたは画像パッチ）と最も関連しているかを捉えることができる．ランキングは，このスコア $s(q, d_j)$ の降順で決定される．

表 1　各手法・条件における nDCG@k（NTCIR-16 Data Search 2）

| 手法 | 条件 | nDCG@1 | nDCG@3 | nDCG@5 | nDCG@10 |
|---|---|---|---|---|---|
| BM25 | メタデータ | 0.112 | 0.108 | 0.119 | 0.116 |
| | メタデータ，統計データ | 0.034 | 0.033 | 0.037 | 0.033 |
| BM25+DA+PRP-Sliding-k | メタデータ，統計データ | 0.212 | 0.222 | **0.260** | **0.284** |
| ColQwen2.5 | メタデータ | **0.259** | **0.241** | 0.244 | 0.230 |
| | メタデータ，統計データ | 0.138 | 0.163 | 0.163 | 0.177 |
| | メタデータ，統計データ（PDF w/metadata） | 0.155 | 0.167 | 0.176 | 0.178 |

## 4 実　　験

### 4.1 実験目的

本実験の目的は，統計データ検索タスクにおいて，ColQwen に基づく視覚的文書検索アプローチの有効性を検証することである．

### 4.2 データセット

実験では，統計文書集合として NTCIR–16 Data Search 2 で提供されている統計データ検索用の英語データセットを利用する [6]．このデータセットには，統計文書集合とテストクエリ集合が含まれている．統計文書の統計データは，多様なデータ形式を含んでいるが，csv と pdf のみを利用した．また，NTCIR–16 上での評価のために，一部の統計文書に対してクエリとの関連性スコア（qrels）が付与されている．関連性スコアは，L0（関連していない），L1（部分的に関連している），L2（関連している）の 3 段階で付与されている．本研究では，これらの適合判定に従って評価を行う．

### 4.3 比較手法

本節では，比較手法を手法ごとに整理する．本実験では，以下の入力データを用いる．「タイトル」および「説明文」は，NTCIR–16 Data Search 2 が各統計文書に付与するメタデータ（$m_j$）中の `title` および `description` を指す．「統計データ」は，統計データ本体（PDF，CSV）を指す．BM25 では，(1) メタデータ（タイトルと説明文），(2) メタデータと統計データ，の 2 条件を比較する．

#### 4.3.1 BM25

本研究の BM25 検索は，純 Python 実装である bm25s を用いて実装した [9]．各条件において，対応するテキストを連結して検索対象とする．

#### 4.3.2 LLM を用いた文書補強とリランキング（BM25+DA+PRP-Sliding-k）

黒川ら [15], [16] は，統計データの数値情報を言語化して検索に用いる手法を提案している．本研究では，その中で最も高い性能が報告されている BM25+DA+PRP-Sliding-k を比較手法として採用する．

この手法は，以下の 2 段階で構成される．第一に，文書補強（Document Augmentation; DA）である．PDF や CSV から抽出した表データの行・列ヘッダと値に基づき，LLM を用いてその内容を説明する自然言語テキストを生成・要約し，これをメタデータと結合して BM25 による初期検索を行う．

第二に，Pairwise Ranking Prompting (PRP) の Sliding-k 手法によるリランキングである．PRP は，2 つの文書とクエリを LLM に入力し，どちらがより関連性が高いかを判定させる手法である．特に Sliding-k 手法は，バブルソートのようにランキングリストの下位から上位に向けてペア比較と並べ替えを繰り返すアプローチであり，限られた計算コストでリスト上位の精度を集中的に改善することを目的としている．この手法は，英語データセットにおいて高い検索精度を示しているが，LLM による推論コストが検索時間に大きく影響する特徴がある．

#### 4.3.3 ColQwen2.5（提案手法）

提案手法である ColQwen2.5 では，各入力データを以下のようにエンコードする．テキスト（タイトル，説明文，CSV）はテキストエンコーダで，PDF は画像エンコーダでページ単位にエンコードする．なお，CSV ファイルはモデルの最大入力長を超えることが多いため，先頭 8192 トークンまでを使用し，超過分は切り捨てた．複数の入力を組み合わせる条件では，それぞれを独立にエンコードし，文書のマルチベクトル表現として用いる．

ColQwen2.5 では，(1) メタデータ（タイトルと説明文），(2) メタデータと統計データ，(3) メタデータと統計データ（PDF w/metadata）の 3 条件を比較する．ColQwen2.5 は単一画像のみを入力とするため，PDF の各ページを独立にエンコードすると文書全体の文脈が失われる．そこで，(3) の条件では，PDF 全体の内容を要約しているメタデータ（タイトル・説明文）のテキストを各ページ画像と同時に VLM へ入力し，ページ単位の埋め込みに文書全体の文脈を付与する．加えて，タイトルと説明文を独立にエンコードしたベクトルも文書のマルチベクトル表現に含める．以降，この条件を「統計データ（PDF w/metadata）」と表記する．

### 4.4 評価指標

検索精度には nDCG@k 用いる．nDCG@k は，検索結果の上位 k 件における適合度を評価する指標であり，上位の結果ほど重要視される．値は 0 から 1 であり，1 に近いほど理想的なランキングを示す．本研究では $k = 1, 3, 5, 10$ の 4 種類で評価する．

### 4.5 実験結果

表 1 に，各手法および各条件における検索精度（nDCG@k）

を示す．

まず，ベースラインである BM25 の結果を述べる．メタデータのみを用いた条件では nDCG@10=0.116 となった．メタデータと統計データを組み合わせた場合は 0.033 となり，統計データを含む条件では精度が著しく低下した．

次に，LLM を用いた既存手法（BM25+DA+PRP-Sliding-k）の結果を述べる．この手法は nDCG@1=0.212，nDCG@3=0.222，nDCG@5=0.260，nDCG@10=0.284 となり，BM25 の全条件を上回る精度を示した．

最後に，提案手法である ColQwen2.5 の結果を述べる．メタデータのみを用いた条件では nDCG@10=0.230 となり，ColQwen2.5 の全条件中で最も高い精度を達成した．メタデータと統計データを組み合わせた場合は nDCG@10=0.177 となり，メタデータのみの条件より低下した．メタデータと PDF（メタデータ付与）を組み合わせた場合は nDCG@10=0.178 となり，統計データを組み合わせた場合と同程度の精度であった．

## 5 考　　察

### 5.1 ColQwen2.5 と BM25 の比較

メタデータのみを用いた条件において，ColQwen2.5（nDCG@10=0.230）は BM25（nDCG@10=0.116）を大きく上回った．

この差は，両手法の検索原理の違いに起因すると考えられる．BM25 は語彙的マッチングに基づき，クエリと文書間でキーワードが一致する場合にのみ高いスコアを与える．一方，ColQwen2.5 は Qwen2.5-VL を基盤としており，大規模コーパスでの事前学習により獲得した言語知識を保持している．これにより，クエリと文書で異なる表現が用いられていても，意味的に近い場合にはマッチングが可能となる．

ColQwen2.5 のマルチベクトル表現では，タイトルと説明文を独立にエンコードするため，それぞれの情報が相互に干渉せず，相補的に機能したと考えられる．

### 5.2 統計データを含む条件の性能低下

BM25 において，統計データを含む条件では精度が著しく低下した（nDCG@10: 0.033）．この原因として，CSV ファイルの構造的特性が挙げられる．CSV の 1 行目や 1 列目にはヘッダー情報（項目名や地域名など）が含まれるが，それ以外のセルは数値の羅列である．BM25 は語彙的マッチングに基づくため，クエリに含まれるキーワードと数値が一致することは稀であり，検索に有効な手がかりとならない．さらに，大量の数値がノイズとして作用し，ヘッダー部分の有用な語彙情報が埋もれてしまったと考えられる．

ColQwen2.5 においても，統計データを含む条件では精度が低下した（nDCG@10: 0.177〜0.178）．この原因として，ページ単位のエンコードでは文書全体の文脈を捉えられない点が挙げられる．例えば，「universities saudi-sourced funding 2012」というクエリに対し，タイトルが「IAWG FY 1999 Annual Report」である文書を考える．この文書のあるページには「universities funding」に関連する内容が含まれているが，それが 2012 年のデータではないことは，タイトルや他のページなど，当該ページ以外の情報を参照しなければ判断できない．しかし，PDF の各ページを独立にエンコードする場合，このような文書全体の文脈情報が失われてしまう．

この問題を解決するため，PDF（メタデータ付与）条件ではメタデータ（タイトル・説明文）を各ページ画像と同時に VLM へ入力することを試みた．しかし，この条件でも精度の改善は限定的であった（nDCG@10: 0.178）．これは，ColQwen2.5 がテキストと画像を同時に入力して文書埋め込みを生成するようには学習されていないことが原因と考えられる．ColQwen2.5 は画像のみ，またはテキストのみを入力とする設定で学習されており，両者を組み合わせた入力に対して最適化されていない．

## 6 おわりに

本研究では，統計データ検索タスクに対して ColQwen2.5 に基づく視覚的文書検索（VDR）を適用し，その有効性と限界を検証した．

実験の結果，メタデータのみを用いた条件において，ColQwen2.5 は BM25 を一貫して上回る精度を示した．これは，ColQwen2.5 が意味的マッチングを可能にすることに加え，マルチベクトル表現により複数の入力を相補的に活用できるためと考えられる．一方，統計データ（PDF, CSV）を含む条件では，BM25・ColQwen2.5 ともに精度が低下した．ColQwen2.5 では，ページ単位のエンコードにより文書全体の文脈が失われることが主な原因であり，メタデータを統合入力する試みも限定的な効果にとどまった．

今後の課題として，以下が挙げられる．第一に，複数ページにまたがる文脈を捉えるための手法の検討である．第二に，VDR モデルの統計データドメインへのファインチューニングの可能性の検討である．

## 謝　　辞

本研究の一部は科研費 23K11342 の助成を受けたものである．

## 文　　献

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[3] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

[4] Jonathan Herzig, Paweł K 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*, pp. 4320–4333, 2020.

[5] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[6] Makoto P. Kato, Ohshima Hiroaki, Liu Ying-Hsang, Chen Hsin-Liang, and Nakano Yu. Overview of the ntcir-16 data search 2 task. p. none, 06 2022.

[7] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

[8] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

[9] Xing Han Lù. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024.

[10] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

[11] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[12] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[13] 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*, pp. 8413–8426, 2020.

[14] 岡本卓, 宮森恒, Okamoto Taku, Miyamori Hisashi. 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）, Vol. 14, No. 4, pp. 36–48, 10 2021.

[15] 黒川博生, 宮森恒, Kurokawa Hiroki, Miyamori Hisashi. Llmを用いた文書補強とリランキングによる広範な統計データ検索. データ工学と情報マネジメントに関するフォーラム（DEIM 2025）, 2025.

[16] 黒川博生, 宮森恒, Kurokawa Hiroki, Miyamori Hisashi. 大規模言語モデルを用いた文書補強とリランキングによる統計データ検索. 情報処理学会論文誌データベース（TOD）, Vol. 18, pp. 20–34, Jul 2025.

# 日本語検索タスクにおける機械翻訳テストコレクションの妥当性検証

岩間　悠莉† 　加藤　誠††,†††

† 筑波大学 知識情報・図書館学類　〒 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
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †yiwama@klis.tsukuba.ac.jp, ††mpkato@acm.org

**あらまし**　本研究では，機械翻訳テストコレクションと，人手で構築されたテストコレクションとを比較して，情報検索システムの性能評価においてどの程度一貫した評価を与えるのか明らかにすることを目的とする．本検証に向けて，英語・日本語間で同一内容であるとみなせるテストコレクションを構築した．このうち英語のテストコレクションを日本語に機械翻訳することで，構築方法のみが異なる同一言語・同一内容の 2 つのテストコレクションを得た．これらを用いて，複数の検索モデルによる検索評価を行い，評価結果の類似性を分析するとともに，異なる翻訳モデルを用いた場合の評価結果についても確認した．実験の結果，機械翻訳テストコレクションの一定の妥当性が示唆された．
**キーワード**　情報検索，評価・データセット，機械翻訳

## 1　はじめに

情報検索システムの性能を客観的に評価するためには，クエリ・文書・適合性判定から構成される評価用データセットであるテストコレクションが不可欠である．しかし，その整備状況には大きな言語間格差があり，英語については多様なテストコレクションが蓄積されている一方で，多くの英語以外の言語では十分なリソースが存在しない．さらに，高品質なテストコレクションの作成には人手による適合性判定が必要であり，そのコストは極めて高い．例えば，18 言語で構築された多言語テストコレクションでは，その構築に 5 人年を要していたと報告されている [22]．このような状況下で，英語以外の言語における情報検索研究や実システムの評価は，英語と比較して利用可能なテストコレクションの制約を受けている．

英語以外の言語でのテストコレクション不足を補う手段として，既存のテストコレクションを機械翻訳して利用する試みが既に行われている．しかし，これらの研究においてはその評価の妥当性が検証されていないことがある [11], [20]．あるいは翻訳前のテストコレクションと翻訳後の他言語テストコレクションとの評価結果を比較し，両者が近いほど望ましいと暗黙に仮定している [6]．

ここで重要となるのは，翻訳後の言語において人手で構築された同一内容のテストコレクションが存在した場合に，機械翻訳テストコレクションが，情報検索システムの性能評価において同様の結論を与えられるかという点である．人手で構築されたテストコレクションは，対象言語における情報検索評価の基準として広く用いられており，一般に実際の情報要求を反映していると考えられている．機械翻訳テストコレクションの妥当性については，このような人手テストコレクションとの比較を通じて，評価結果がどの程度整合するかという観点から検討することが，1 つの有効なアプローチである．ここで「同一内容」とは，クエリが同一の情報要求を表し，文書および適合性判定についても同一の情報を伝えているとみなせる設定を意味する．また，本研究において「同様の評価」とは，評価値の絶対的な一致を直接比較するのではなく，検索モデル間の相対的な性能比較やモデル順位，文書ランキングの傾向といった評価結果の構造に着目した比較を指す．しかし，既存研究ではこのような観点からの直接的な比較は行われておらず，機械翻訳テストコレクションによる評価の妥当性は十分に検証されていない．

本研究の目的は，機械翻訳テストコレクションと，人手で構築された同一言語・同一内容のテストコレクションと比較して，情報検索システムの性能評価においてどの程度一貫した結論を与えるのかを定量的に明らかにすることである．そのために本研究では，英語と日本語の 2 言語に焦点を当て，同一内容を表すクエリおよび文書が言語間で対応付けられたクエリ集合・文書集合を組み合わせることで，テストコレクションを構築する．次に，このうち英語のクエリと文書を日本語へ機械翻訳することで「人手で作成された日本語テストコレクション」と「英語から機械翻訳された日本語テストコレクション」という 2 種類の日本語テストコレクションが得られる．これらは構築方法のみが異なる，同一言語・同一内容のテストコレクションとみなせる．これを用いて，複数の検索モデルによる評価実験を行い，人手テストコレクションと機械翻訳テストコレクションそれぞれに対して得られる評価値や検索モデルの順位を比較することで，機械翻訳テストコレクションによる評価の妥当性を検証する．

本研究では，実験用のテストコレクションとして，多言語質問応答データセットである MKQA [10] を Wikipedia と組み合わせることで構築した英語・日本語間で同一内容とみなせるテストコレクションを用いた．評価実験では疎検索モデルである BM25 といくつかの密検索モデルを対象とし，人手テストコレクションと機械翻訳テストコレクションの評価結果を比較した．比較にあたっては，検索モデルの順位の一致度を

Kendall's $\tau$ により，各クエリにおける文書ランキングの一致度を Rank-Biased Overlap により評価した．さらに，翻訳品質の影響を分析するため，Google 翻訳と 6 種類の公開翻訳モデルを用い，翻訳品質については BLEU により概算した．

本研究の実験結果から，機械翻訳テストコレクションは，検索モデルの順位の類似度という観点では，人手で構築されたテストコレクションと概ね一貫した結論を与えることが確認された．一方で，翻訳品質の違いが評価結果に与える影響については，極端に翻訳品質が低い場合を除き，翻訳品質の高さが妥当性の向上に直結するとは限らないことが示された．

本研究における貢献を以下に示す：

1. 英語・日本語間で対応するテストコレクションを用い，同一言語・同一内容の人手テストコレクションと機械翻訳テストコレクションを比較することで，機械翻訳テストコレクションによる評価の妥当性を検証した．
2. MKQA と Wikipedia を組み合わせることで，情報検索評価に利用可能な英語・日本語間で同一内容とみなせるテストコレクションを構築し，本研究の実験で利用した．
3. Google 翻訳と 6 種類の公開翻訳モデルによる機械翻訳結果を用いて，翻訳品質と機械翻訳テストコレクションによる評価の妥当性との関係を分析した．

本論文の構成は次の通りである． 第 2 節では，本研究と関連する既存研究について述べる．第 3 節では，構築したテストコレクションおよび実験設定について説明する． 第 4 節では，実験結果を示し，機械翻訳テストコレクションの妥当性について考察する．第 5 節では，今後の課題と共に本研究の結論を述べる．

## 2 関 連 研 究

日本語の情報検索評価のためのテストコレクションとしては，これまでにいくつかの取り組みが存在する．例えば，日本語を対象とした情報検索評価の枠組みとして，国立情報学研究所が主催する NTCIR プロジェクトにおいて，これまでに複数のテストコレクションが構築されてきた [7]．一方で，NTCIR におけるテストコレクションは，タスクや対象ドメインが多様であるが，利用にあたっては申請手続きが必要であり，研究目的での利用が前提とされている．また，多言語を対象とした情報検索評価用テストコレクションにおいても，日本語を対象言語の一つとして含むものが提案されている．代表的な例として，MIRACL [22] や Mr.TyDi [21] が挙げられる．近年では，日本語のテキスト埋め込みの評価を目的としたベンチマークとして，JMTEB[1]が公開されている．このベンチマークは，検索を含む複数の評価タスクを対象としており，その一部として，上述の MIRACL や Mr.TyDi を含む既存の検索評価用テストコレクションが収録されている．一方で，JMTEB に含まれるその他のテストコレクションには，クエリの生成方法や適合性判定の方法といった点で多様な背景を持つため，各テストコレクションの特性を考慮することが重要である．このように，日本語においては，英語における BEIR [16] に代表されるような，広く利用される汎用的な情報検索ベンチマークは十分に整備されていない．

こうした状況は日本語に限ったものではなく，英語以外の言語においても高品質なテストコレクションの不足が指摘されている [11], [20]．この課題に対する 1 つのアプローチとして，既存の英語テストコレクションを他言語へ機械翻訳して利用する試みが行われている．Wojtasik らは，英語における大規模な情報検索ベンチマークである BEIR をポーランド語に機械翻訳することで，ポーランド語における情報検索評価のためのベンチマーク BEIR-PL を構築した [20]．同研究では，翻訳後のテストコレクションを用いて多数の検索モデルを対象とした評価実験を行い，モデル性能に関する結果を報告している．一方で，翻訳後のテストコレクションによる評価の妥当性については明示的な検証は行われていない．同様に Lotfi らは，BEIR をオランダ語に機械翻訳することで，オランダ語におけるベンチマーク BEIR-NL を構築した [11]．同研究では，翻訳後のテストコレクションを用いて多数の検索モデルを対象とした評価実験の結果を報告している．また，一部のモデルを用いた BEIR や BEIR-PL による評価結果との比較や，BEIR-NL を英語に逆翻訳したベンチマークと元の英語 BEIR との比較により，機械翻訳という構築方法の評価結果への影響を部分的に検証している．しかし，この分析は，機械翻訳後のテストコレクションが，同一言語における人手のテストコレクションと同様の評価を行えるかを検証したものではない．また，Jeronymo らは，代表的な情報検索評価ワークショップである TREC において構築された英語のアドホック検索用テストコレクションである Robust04 を機械翻訳することで，mRobust04 を構築した [6]．同研究でも，翻訳後のテストコレクションを用いて複数の検索モデルを対象とした評価実験の結果を報告している．ただし，その評価の妥当性についての明示的な検証は行われていない．

テストコレクションが十分に存在しない状況に対する別のアプローチとして，大規模言語モデルを用いて適合性判定を行う手法が提案されている．いわゆる LLM-as-a-Judge は，クエリと検索結果文書を入力として，大規模言語モデルによる適合性判定を行うことで，人手による適合性判定を代替することを目的としている [4], [23]．加えて，大規模言語モデルを用いてクエリおよび適合性判定結果を事前に生成し，人手に依らないテストコレクションを構築する手法として，合成テストコレクションが提案されている．Rahmani らは，構築した合成テストコレクションと人手により構築されたテストコレクションを比較し，評価結果の類似性を分析している [14]．こうした手法は，人手による適合性判定を用いずに評価を行える利点を持つ一方で，生成手法や用いるモデルに依存した特性を持つ可能性があり，評価結果に影響を与え得る体系的なバイアスが生じることも報告されている [15]．

表 1 に示すように，テストコレクション不足への対処法には，機械翻訳テストコレクション，LLM-as-a-Judge，合成テストコレクションといった手法が存在し，それぞれ前提とするデータや適合性判定の生成方法において異なる特性を持つ．

1：https://huggingface.co/datasets/sbintuitions/JMTEB

表 1: テストコレクション不足への対処法の比較

| 手法 | 前提とするデータ | 適合性判定 |
|---|---|---|
| 機械翻訳テストコレクション | **英語** 文書・クエリ・適合性判定 | 人手 |
| LLM-as-a-Judge | **日本語** 文書・クエリ | LLM |
| 合成テストコレクション | **日本語** 文書 | LLM |

LLM-as-a-Judge や合成テストコレクションについては，人手により構築されたテストコレクションと類似した評価結果が得られることが報告されている一方で，生成手法や用いるモデルに依存する可能性も指摘されており，その妥当性については引き続き検討が行われている．

機械翻訳テストコレクションは，既存のテストコレクションに含まれるクエリおよび文書を機械翻訳によって他言語へ変換し，翻訳前のテストコレクションにおける適合性判定との対応付けを保ったまま構築されるテストコレクションである．このような方法は，人手による適合性判定を再利用できるという実用的な利点を持つ一方で，機械翻訳を介した場合に評価結果がどの程度変化するのかについては，十分に明らかになっていない．本研究では，この機械翻訳テストコレクションに着目し，人手により構築された同言語・同一内容のテストコレクションとの比較を通じて，評価結果の関係性を分析する．

## 3 実験設定

本節では，本研究における実験設定について述べる．まず，日英間で同一の内容を持つテストコレクションの構築の概略について説明する．次に，機械翻訳によるテストコレクションの生成の概要を述べる．続いて，本研究で使用した検索モデル，評価方法について述べる．

### 3.1 英語・日本語間で対応するテストコレクションの構築

英語・日本語間で対応する情報検索テストコレクションは，既存には公開されていない．そのため，機械翻訳テストコレクションと人手で構築されたテストコレクションを同一内容の条件下で比較することは困難である．本研究ではこの課題に対処するため，英語と日本語の間で対応付けられたクエリ集合，文書集合，および適合性判定からなる本研究の目的に即したテストコレクションを構築した．

本研究では，異なる言語間で同一の内容を表すクエリ集合を得るため，多言語質問応答データセットである MKQA [10] を用いた．MKQA は，英語の質問応答データセットである Natural Questions [8] を起点として，その英語の質問文を人手により多言語へ翻訳することで構築されたデータセットであり，英語と日本語を含む 26 言語に対応している．本研究では，このうち英語および日本語の質問文を，テストコレクションにおけるクエリとして利用した．

文書集合には Wikipedia を用いた．これは，MKQA の元となっている Natural Questions が Wikipedia の記事を情報源として構築されているためであり，質問応答データを情報検索タスクへ変換する際に一般的に用いられている設計と整合している．本研究では，英語版および日本語版 Wikipedia の記事データを対象とし，2025 年 7 月 20 日版のダンプを利用した．

表 2: 構築したテストコレクションの統計

| 言語 | クエリ数 | 文書数 | 適合性判定数 |
|---|---|---|---|
| 英語 | 1,263 | 663,609 | 1,263 |
| 日本語 | 1,263 | 669,824 | 1,263 |

文書集合の対応付けには，Wikipedia における言語間リンクを用いた．言語間リンクは，異なる言語版における同一概念の記事同士を結び付けるものである．抽出前の文書数は，英語版 Wikipedia で 7,009,646 件，日本語版 Wikipedia で 1,459,624 件であった．これらの中から，英語版の記事のうち日本語記事への言語間リンクを持つもの，および日本語版のうち英語版への言語間リンクを持つものをそれぞれ抽出した．その結果，英語では 663,609 件，日本語では 669,824 件の文書が得られた．Wikipedia の記事は，言語間リンクで結ばれていても，内容が必ずしも一致するとは限らない．そこで本研究では，文書内容の差異による影響を緩和するため，各記事の冒頭のセクションのみを利用した．冒頭のセクションは記事全体の概要を含むことが多く，全文を用いる場合と比較して，内容の差異が小さいと期待され，言語間の内容差による影響を抑制できると考えられる．次に，文書集合への制約に伴い，利用するクエリおよび適合性判定結果を限定した．MKQA には 10,000 件のクエリと適合性判定結果が含まれているが，言語間リンクにより結ばれた文書中に適合文書が存在するクエリのみを抽出した．結果として，5,221 件のクエリが得られた．加えて，文書を冒頭セクションに限定したことにより，その文書を適合文書と判断する根拠となる情報が失われる影響を抑えるため，MKQA における解答が文書中に直接含まれる場合のみを適合文書として採用した．この条件で適合文書をもつクエリは 1,263 件であった．以上の手順より，日英で対応するクエリ集合，文書集合および適合性判定結果からなるテストコレクションを構築した．

表 2 に，構築したテストコレクションの統計を示す．文書集合のサイズは言語間で完全には一致していないが，後述する機械翻訳テストコレクションの生成および検索評価では，適合文書と同一の手続きで抽出したランダム文書を用いて文書集合を構成しており，この不均衡は評価結果に影響しない設計となっている．

### 3.2 機械翻訳テストコレクションの生成

本節では，3.1 節で構築した英語テストコレクションを機械翻訳することで，機械翻訳による日本語テストコレクションを生成する手順について述べる．翻訳対象としたのは，英語テストコレクションに含まれる全てのクエリ，それらに対応する適合文書，および検索評価に用いるためにランダムに抽出した 50,000 件の文書である．適合文書に加えてランダム文書を翻訳対象に含めることで，全文書を翻訳することなく，検索評価に必要な数の非適合文書を含む文書集合を構成した．その結果，機械翻訳テストコレクションは，1,263 件のクエリ，51,263 件

表 3: Tatoeba データセットにおける翻訳品質評価

| 翻訳モデル | BLEU |
|---|---|
| Google 翻訳 | 30.90 |
| M2M-100 (1.2B) | 21.13 |
| NLLB-200 (3.3B) | 20.14 |
| mBART-50 | 19.90 |
| M2M-100 (418M) | 18.85 |
| NLLB-200 (600M) | 17.35 |
| OPUS-MT | 0.88 |

の文書，1,263 件の適合性判定から構成されており，人手テストコレクションと対応関係を持つ．

機械翻訳には，Google 翻訳 [2]，NLLB-200 (3.3B[3]，600M[4]) [3]，M2M-100 (1.2B[5]，418M[6]) [5]，mBART-50[7] [9]，および OPUS-MT[8] [17], [18] を用いた．Google 翻訳は，既存の機械翻訳による情報検索データセットの構築において広く利用されている翻訳サービスであり [2], [6], [20]，本研究においても代表的な翻訳手法として採用した．加えて，翻訳モデルの違いが検索評価に与える影響を確認するため，多言語翻訳を目的として学習された複数の公開モデルを用いた．

### 3.3 翻訳モデルの品質評価

本研究では，機械翻訳テストコレクションの生成に複数の翻訳モデルを用いているため，検索評価実験に先立ち，各翻訳モデルの翻訳品質を自動評価指標により比較した．翻訳品質の評価には，mMARCO [2] における翻訳品質評価の設定に倣い，Tatoeba データセット [9]を用いて BLEU スコア [12] を算出した．具体的には，Tatoeba データセットの 2023 年 4 月 12 日版に含まれる英語と日本語の翻訳ペアの test 分割から，1,000 件をランダムに抽出し，各翻訳モデルによる翻訳結果と参照訳との間で BLEU スコアを計算した．BLEU の算出には SacreBLEU [13] を用い，日本語のトークナイザとして ja-mecab を指定した．

その結果は表 3 に示す通りであり，Google 翻訳が最も高い BLEU スコアを示した．次いで，M2M-100 (1.2B)，NLLB-200 (3.3B)，mBART-50，M2M-100 (418M)，NLLB-200 (600M) の順となった．一方，OPUS-MT は他のモデルと比較して極端に低い BLEU スコアを示した．

### 3.4 検索モデル

本節では，構築・生成した各テストコレクションに対して検索評価を行うために用いた検索モデルについて述べる．いずれの検索モデルも事前学習済みモデルをそのまま使用し，追加の学習やチューニングは行っていない．

2：https://cloud.google.com/translate
3：https://huggingface.co/facebook/nllb-200-3.3B
4：https://huggingface.co/facebook/nllb-200-distilled-600M
5：https://huggingface.co/facebook/m2m100_1.2B
6：https://huggingface.co/facebook/m2m100_418M
7：https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
8：https://huggingface.co/Helsinki-NLP/opus-mt-en-jap
9：https://tatoeba.org

a） 疎検索モデル

疎検索モデルとして，BM25 を用いた．BM25 は，単語の出現頻度に基づく語彙一致型の情報検索モデルであり，情報検索分野において広く利用されている．実装には Pyserini を用い，Lucene に基づく BM25 を使用した．ハイパーパラメータは Pyserini のデフォルト設定に従い，$k_1 = 0.9$，$b = 0.4$ とした．

b） 密検索モデル

密検索モデルとして，多言語情報検索において近年広く利用されている事前学習済み埋め込みモデルを複数用いた．使用したモデルは，mDPR，mContriever，LaBSE，mE5 (small, base, large)，mGTE，jina-embeddings-v3，bge-m3 である．実装については，mDPR は MIRACL [22] などで用いられている MS MARCO [1] によって事前学習されたモデルを用いた．その他のモデルには Sentence-Transformers を利用した．いずれのモデルについても，事前学習済みモデルをそのまま使用してゼロショットで検索評価を行った．

### 3.5 評価方法

本節では，人手で構築されたテストコレクションと機械翻訳テストコレクションにおける検索評価結果を比較するために用いた評価方法について述べる．

#### 3.5.1 検索性能の評価指標

検索結果の性能を評価する指標として，Recall@100 と nDCG@10 を用いた．Recall@100 は，上位 100 件の検索結果に適合文書が含まれるかどうかを評価する指標であり，適合文書を検索結果中に発見できたかどうかを測るために用いた．nDCG@10 は，検索結果上位において適合文書がどの位置に現れるかを考慮した評価指標である．本研究で用いたテストコレクションでは，各クエリに対する適合文書は高々1 件であり，適合性は 2 値で与えられているため，nDCG@10 は検索結果上位 10 件以内に適合文書が出現するかどうか，およびその順位を反映した指標となる．

#### 3.5.2 評価結果の類似度指標

人手のテストコレクションと機械翻訳テストコレクションに基づく検索評価結果の類似度を分析するため，本研究では順位付けの類似度を測る指標を用いた．評価結果の比較は，検索モデルの順位と，各クエリにおける文書ランキングという 2 つの異なる粒度で行う．

a） 検索モデルの順位の類似度

検索モデルの順位の類似度を評価するため，Kendall's $\tau$ を用いた．Kendall's $\tau$ は，2 つの順位付けの間の相関を測る指標であり，順位の全体的な一致度を評価できる．本研究では，各テストコレクションにおいて得られた検索モデルの評価値に基づき，モデル順位を作成し，人手のテストコレクションと機械翻訳テストコレクションとの間での Kendall's $\tau$ を算出した．なお，Kendall's $\tau$ により得られた順位相関の統計的有意性を検討するため，統計的検定を行った．

b） 文書ランキングの類似度

一方，各クエリにおける文書ランキングの類似度を評価するため，Rank-Biased Overlap (RBO) [19] を用いた．RBO は，

2 つの文書ランキング $S$ および $T$ に対して，ランキングの深さ $d$ を順に増やしながら，上位 $d$ 件に含まれる文書の一致度を用いて類似度を評価する指標である．文書ランキングの順序関係を一様に評価する Kendall's $\tau$ とは異なり，検索結果として重要な上位順位の一致を重視できる点に特徴がある．RBO は次式で定義される：

$$\mathrm{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} A_d, \tag{1}$$

$$A_d = \frac{|S_{1..d} \cap T_{1..d}|}{d}, \tag{2}$$

ここで $S_{1..d}$ および $T_{1..d}$ は，それぞれ文書ランキング $S$ および $T$ の上位 $d$ 件を表す．$A_d$ は，順位 $d$ までに含まれる文書集合の一致度，すなわち上位 $d$ 件に含まれる共通文書の割合を意味する．$p$ は上位順位をどの程度重視するかを制御するパラメータであり，本研究では RBStar[10]のデフォルトのパラメータである $p = 0.95$ を用いた．

本研究では，各クエリについて上位 1,000 位までの文書ランキングを得ているが，上位の順位が検索結果として特に重要であることから，RBO を用いて文書ランキングの類似度を分析した．

## 4 実験結果

本節では，本研究で実施した検索評価実験の結果を示す．機械翻訳テストコレクションの妥当性を検証するため，本節では 2 つの観点から分析を行う．まず，人手で構築されたテストコレクションと機械翻訳テストコレクションに基づく検索評価結果を比較し，検索モデルの順位の一致度に基づいて妥当性を分析する．次に，翻訳モデルごとの評価結果を比較し，翻訳品質が評価結果に与える影響を分析する．

### 4.1 検索モデルの順位の一致度に基づく妥当性分析

本節では，機械翻訳テストコレクションに基づく検索評価結果が，人手で構築されたテストコレクションによる評価結果と同様の検索モデルの順位を与えるかを分析する．

まず，人手テストコレクションと機械翻訳テストコレクションに基づく検索評価値の対応関係を確認するため，散布図による比較を行った．図 1a および図 1b は，それぞれ Recall@100 および nDCG@10 における，各検索モデルの評価値を，人手テストコレクション（横軸）と機械翻訳テストコレクション（縦軸）で対応付けたものである．これらの散布図から，OPUS-MT を除く多くの翻訳モデルにおいては，機械翻訳テストコレクションに基づく評価値が，人手テストコレクションによる評価値と概ね単調な関係を保っていることが確認できる．一方で，OPUS-MT による機械翻訳テストコレクションでは，検索性能の評価結果が著しく低下しており，他の翻訳モデルとは異なる傾向を示す．

次に，評価値そのものではなく，検索モデル間の相対的な順位関係に着目し，人手テストコレクションと機械翻訳テストコレクションに基づく評価結果の一致度を Kendall's $\tau$ により定量的に評価した．表 4 は，各翻訳モデルを用いた機械翻訳テストコレクションと人手テストコレクションとの間で算出した Kendall's $\tau$ を示している．その結果，OPUS-MT を除く全ての翻訳モデルにおいて，Recall@100，nDCG@10 のどちらの指標においても概ね高い Kendall's $\tau$ の値が得られた．また，これらの順位相関はいずれも有意水準 0.05 において有意であり，人手テストコレクションと機械翻訳テストコレクションの間に統計的に有意な相関が認められた．一方，OPUS-MT については，Kendall's $\tau$ の値が低く，統計的に有意な相関は認められなかった．

表 4: 人手テストコレクションと機械翻訳テストコレクション間のモデル順位の Kendall's $\tau$

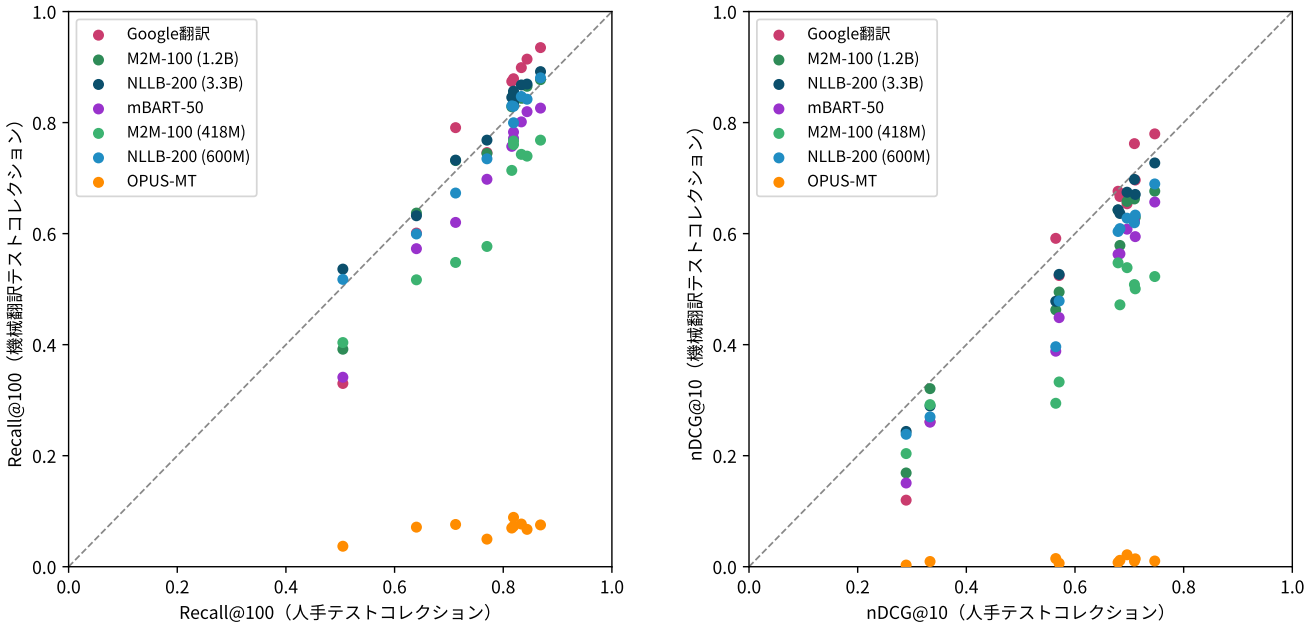
| 翻訳モデル | Recall@100 | nDCG@10 |
|---|---|---|
| Google 翻訳 | 0.8989 | 0.7778 |
| M2M-100 (1.2B) | 0.9439 | 0.8222 |
| NLLB-200 (3.3B) | 0.9439 | 0.8667 |
| mBART-50 | 0.9888 | 0.9111 |
| M2M-100 (418M) | 0.7641 | 0.6000 |
| NLLB-200 (600M) | 0.8540 | 0.9556 |
| OPUS-MT | 0.3146 | 0.3778 |

なお，テストコレクションの代替可能性を検索システム順位の Kendall's $\tau$ により評価した先行研究として，Rahmani らは，クエリおよび適合性判定の双方を LLM により合成したテストコレクションと人手テストコレクションとの間で，nDCG@10 において $\tau = 0.8568$ の一致度を示している [14]．また，Faggioli らは，適合性判定のみを LLM により代替した設定において $\tau = 0.86$ の一致が得られることを示している [4]．これらの研究とは実験設定が異なるものの，OPUS-MT を除く本研究で得られた Kendall's $\tau$ の値は，先行研究において高い一致度が示されている水準と同程度であると言える．以上より，本研究の実験設定においては，一定の翻訳品質を持つ翻訳モデルを用いた場合，機械翻訳テストコレクションは検索モデルの順位の評価という観点から，人手テストコレクションの代替として妥当な評価結果を与えることが示唆された．

### 4.2 翻訳品質の違いが評価結果に与える影響

前節では，検索モデルの順位の一致度に着目し，機械翻訳テストコレクションが人手テストコレクションと高い一致度を示すことを確認した．一方，機械翻訳テストコレクションは翻訳モデルに依存して構築されるため，翻訳モデルの違いが評価結果にどのような影響を与えるかを，より詳細に分析する必要がある．表 3 で示した翻訳品質評価では，Google 翻訳，M2M-100 (1.2B)，NLLB-200 (3.3B)，mBART-50，M2M-100 (418M)，NLLB-200 (600M)，OPUS-MT の順で BLEU スコアが低下することが確認された．本節では，この翻訳品質の違いが検索評価結果にどのような影響を与えるかを分析する．

10：https://github.com/rankbiased/rbstar




(a) Recall@100　　(b) nDCG@10

図 1: 機械翻訳テストコレクションと人手テストコレクションによる検索性能評価結果

表 4 で示した翻訳モデル別の Kendall's $\tau$ に着目すると，OPUS-MT を除くいずれの翻訳モデルにおいても，検索モデルの順位の一致度は高い水準にあることが分かる．しかし，翻訳モデル間で Kendall's $\tau$ の値には差が見られ，その順序関係は翻訳品質評価の結果とは必ずしも一致しない．例えば，BLEU スコアが最も高い Google 翻訳よりも，mBART-50 の方が高い Kendall's $\tau$ を示している．このことから，翻訳品質の自動評価指標による優劣が，検索評価結果の一致度に単純には反映されないことが示唆される．

もっとも，Kendall's $\tau$ は検索モデルの順位関係という比較的粗い粒度での一致度を評価する指標であり，翻訳モデルの違いが検索評価結果に与える影響を十分に捉えきれていない可能性がある．そこで次に，各クエリにおいて検索により得られた文書ランキングの一致度に着目し，Rank-Biased Overlap (RBO) を用いた分析を行う．

表 5 は，翻訳モデル別に，各検索モデルにおけるクエリごとの文書ランキングの RBO の平均を示したものである．全ての検索モデルにおいて，Google 翻訳が最も高く，OPUS-MT が最も低い RBO を示している．一方，その他の中間的な翻訳品質を持つ翻訳モデル間では，RBO の順序が BLEU スコアの順序と必ずしも一致していない．さらに，各検索モデルごとに，翻訳モデルを要因とした Tukey HSD 検定を行った結果，有意水準 0.05 において，OPUS-MT は全ての検索モデルにおいて全翻訳モデルとの間に有意差が認められた．Google 翻訳についても同様に，他の全翻訳モデルとの間に有意差が認められた．一方，その他の中間的な翻訳品質を持つ翻訳モデル同士では，有意差が認められないペアが複数存在した．

BLEU スコアと RBO の関係をより直接的に確認するため，各翻訳モデルの BLEU スコアと全検索モデルにおける RBO 平均の間の Pearson の相関係数 $r$ を算出した．全 7 モデルを対象とした場合には $r = 0.953$ と強い正の相関が認められたが，OPUS-MT を除外した場合には $r = 0.794$ となり，有意水準 0.05 において有意な相関は認められなかった．

これらの結果から，極端に翻訳品質が低い場合には RBO が著しく低下する一方，一定の翻訳品質を持つ翻訳モデル間では，BLEU スコアの高さが RBO の向上に直結するとは限らないことが示された．

#### 4.2.1　低い RBO を示すクエリに関する定性的分析

本節では，RBO が特に低いクエリについて行った定性的な観察の結果を述べる．なお，ここでは翻訳モデル間の比較を目的とするものではなく，BM25 および mE5-large の両検索モデルにおいて，Google 翻訳を用いた場合に RBO が下位に位置したクエリを対象とした．

これらのクエリを確認した結果，多くのクエリにおいて人名や作品名などの固有名詞が含まれており，クエリと文書の間で当該固有名詞の翻訳結果が一致していない例が観察された．例えば，楽曲 Seasons in the Sun に関するクエリでは，クエリ中では「太陽の季節」と翻訳された一方，適合文書中では「シーズンズ・イン・ザ・サン」という表記が用いられていた．このように，検索タスクにおいて重要な手がかりとなる固有名詞について，クエリと文書で使用される語彙が一致していない場合，両者の対応関係が弱まり，文書ランキングの一致度が低下する可能性がある．また，固有名詞を含む英語表現が，翻訳過程において固有名詞として適切に解釈されない場合，一般的な語句として処理され，結果として不自然な文法構造の翻訳文が生成されている例も確認された．このような場合には，クエリが検索に適した表現を十分に保持できず，文書側の表現との対応関係がさらに不安定になると考えられる．これらの事例は，機械

表 5: クエリごとの文書ランキングの一致度（RBO）の平均

| 検索モデル | Google | NLLB-200 (3.3B) | M2M-100 (1.2B) | mBART-50 | M2M-100 (418M) | NLLB-200 (600M) | OPUS-MT |
|---|---|---|---|---|---|---|---|
| BM25 | 0.2053 | 0.1627 | 0.1440 | 0.1340 | 0.1025 | 0.1329 | 0.0078 |
| mContriever | 0.1119 | 0.0990 | 0.0759 | 0.0922 | 0.0784 | 0.0887 | 0.0022 |
| LaBSE | 0.1393 | 0.1315 | 0.1232 | 0.1178 | 0.1030 | 0.1159 | 0.0060 |
| mDPR | 0.2471 | 0.2124 | 0.1843 | 0.1802 | 0.1377 | 0.1852 | 0.0041 |
| mE5-small | 0.2772 | 0.2345 | 0.1887 | 0.2024 | 0.1591 | 0.2114 | 0.0048 |
| mE5-base | 0.2865 | 0.2518 | 0.2126 | 0.2231 | 0.1754 | 0.2327 | 0.0065 |
| mE5-large | 0.3460 | 0.2982 | 0.2519 | 0.2666 | 0.2032 | 0.2703 | 0.0059 |
| mGTE | 0.3366 | 0.2936 | 0.2649 | 0.2648 | 0.2099 | 0.2686 | 0.0093 |
| jina-embeddings-v3 | 0.3544 | 0.2930 | 0.2726 | 0.2518 | 0.2150 | 0.2650 | 0.0078 |
| bge-m3 | 0.3150 | 0.2616 | 0.2299 | 0.2341 | 0.1677 | 0.2268 | 0.0052 |

翻訳による意味的な正確さとは独立に，クエリと文書の間で語彙や表記の一貫性が保持されない場合，文書ランキングの一致度が低下し得ることを示している．したがって，RBO に基づく評価結果を解釈する際には，このような翻訳上の特性を考慮する必要がある．

以上の分析から，一定の翻訳品質を持つ翻訳モデルを用いた場合，機械翻訳テストコレクションは検索モデルの順位という観点では人手テストコレクションと概ね一貫した評価結果を与えることが確認された．一方で，翻訳品質の高さが評価結果の妥当性に直結するとは限らず，Kendall's $\tau$ および RBO のいずれにおいても，翻訳品質の自動評価指標の順序と評価結果の一致度の間に単純な対応関係は認められなかった．

## 5 結　　論

本研究では，日本語検索タスクにおける機械翻訳テストコレクションの妥当性を検証することを目的とし，人手で構築されたテストコレクションと機械翻訳テストコレクションに基づく検索評価結果を比較した．

まず，検索モデルの順位の一致度に基づく分析から，極端に翻訳品質が低い OPUS-MT を除き，機械翻訳テストコレクションは人手テストコレクションとの Kendall's $\tau$ において高い値を示すことが確認された．この結果は，一定の翻訳品質を持つ翻訳モデルを用いた場合，検索モデル間の相対的な性能関係という観点では，人手テストコレクションと概ね整合する評価結果が得られることを示している．

一方で，翻訳品質の高さが評価結果の妥当性に直結するとは限らないことも明らかとなった．Kendall's $\tau$ および RBO のいずれにおいても，翻訳品質の自動評価指標の順序と評価結果の一致度の間に単純な対応関係は認められなかった．

本研究で扱った機械翻訳テストコレクションは日本語のものに限られており，他言語においても同様の傾向が見られるかは明らかではない．また，異なる領域の文書や複数の適合文書を含む設定において，機械翻訳モデルがどのように影響するかも未知数である．本研究で用いたテストコレクションの限界を踏まえ，多言語・多領域での検証が今後の課題として挙げられる．

## 謝　　辞

本研究は JSPS 科研費 JP23K28090 の助成，および情報・システム研究機構 “戦略的研究プロジェクト” の支援を受けたものです．ここに記して謝意を表します．

## 文　　献

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2018.

[2] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

[3] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. No language left behind: Scaling human-centered machine translation. *Nature*, Vol. 630, pp. 841–846, 2024.

[4] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 39–50, 2023.

[5] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, Vol. 22, pp. 1–48, 2021.

[6] Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. mRobust04: A multilingual ver-

sion of the TREC Robust 2004 benchmark. *arXiv preprint arXiv:2209.13738*, 2022.

[7] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 11–44, 1999.

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 452–466, 2019.

[9] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 726–742, 2020.

[10] Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open-domain question answering. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1389–1406, 2021.

[11] Ehsan Lotfi, Nikolay Banar, and Walter Daelemans. BEIR-NL: Zero-shot information retrieval benchmark for the Dutch language. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pp. 36–45, 2025.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.

[13] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018.

[14] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. Synthetic test collections for retrieval evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2647–2651, 2024.

[15] Hossein A. Rahmani, Varsha Ramineni, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. Towards understanding bias in synthetic data for evaluation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5166–5170, 2025.

[16] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[17] Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, Vol. 58, pp. 713–755, 2023.

[18] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, 2020.

[19] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, Vol. 28, No. 4, pp. 1–38, 2010.

[20] Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. BEIR-PL: Zero-shot information retrieval benchmark for the Polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2149–2160, 2024.

[21] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multilingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 127–137, 2021.

[22] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, Vol. 11, pp. 1114–1131, 2023.

[23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

# 事前学習済みBERTモデル検索タスクのための評価データセット

ファムフーロン† 三林 亮太†† 荘司 慶行††† 加藤 誠†††,††††† 山本 岳洋†
山本 祐輔†††††† 大島 裕明†

† 兵庫県立大学 情報科学研究科 〒651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 神戸大学 国際文化学研究科 〒651–8501 兵庫県神戸市灘区鶴甲 1–2–1
††† 静岡大学 情報学部 〒432–8011 静岡県浜松市中央区城北 3–5–1
†††† 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
††††† 国立情報学研究所 情報社会相関研究系 〒101–8430 東京都千代田区一ツ橋 2–1–2
†††††† 名古屋市立大学 データサイエンス研究科 〒467–8501 愛知県名古屋市瑞穂区瑞穂町字山の畑 1
E-mail: †{af23a009@guh.u-hyogo.ac.jp, t.yamamoto@sis.u-hyogo.ac.jp, ohshima@ai.u-hyogo.ac.jp},
††mibayashi@people.kobe-u.ac.jp, †††shojiy@inf.shizuoka.ac.jp, {††††,†††††}mpkato@acm.org,
†††††yusuke_yamamoto@acm.org

**あらまし** 本研究では，事前学習済みBERTモデル検索タスクのための評価データセットを構築する．自然言語処理では，事前学習済みモデルをタスクに合わせてファインチューニングして利用することが一般的であるが，多数の候補から目的タスクに適したモデルを選択することは容易ではない．モデルの適性は実際にファインチューニングを行うまで判別しにくく，すべての候補を試行するには多大な時間と計算資源を要するため，タスクに適したモデルを効率的に探索できる仕組みが不可欠である．そこで本研究では，48件の文書分類タスクと20種類のBERTモデルの組合せからなる評価データセットを作成した．さらに，既存手法をベースラインとして実装するとともに，タスクと検索対象モデルの特徴を用いた検索手法を提案し，構築したデータセットを用いてその有効性を検証した．

**キーワード** BERTモデル検索，言語モデル，検索モデル，テキスト分類

## 1 はじめに

事前学習済みBERTモデルは，文書分類や質問応答など，多岐にわたる自然言語処理タスクで用いられている．新たなタスクに取り組む際，ゼロからモデルを学習する代わりに，既存の事前学習済みモデルをファインチューニングして用いることが一般的である．しかし，特定のタスクに最も適した事前学習済みモデルを選択することは容易ではない．

通常，特定のタスクに対する事前学習済みモデルの適合性は，ファインチューニングを通じて評価される．一般的なモデル選定のアプローチは以下の通りである．

1. 多数の事前学習済みモデルの中からいくつかの候補を選択する
2. 1で選択したモデルをタスクの一部のデータ（訓練データ等）を用いてファインチューニングを行う
3. タスクの残りのデータ（テストデータ等）を用いて，2で学習したモデルの性能を評価する
4. 最も性能が良いモデルを採用する

このプロセスにおける最大の問題は，ファインチューニングに要する計算コストの高さである．Hugging Face Hubでは2025年12月24日時点で46,337の事前学習済みBERTモデルが公開されており，これら全てのモデルに対してファインチューニングを行い比較することは不可能である．そのため，実際に学習を行うことなく，事前学習済みBERTモデルを効率的に検索・選択する手法が必要とされている．

事前学習済みモデルの検索あるいは選択に関する問題は，近年活発に研究されており，*Transferability Estimation*などの名称で呼ばれている．しかし，現時点では研究分野として十分に体系化されておらず，問題に対する統一的な定義もなされていない．また，手法の性能を定量的に比較するための標準的なベンチマークデータセットも公開されていないのが現状である．

そこで本研究では，事前学習済みBERTモデル検索という問題を体系的に定義する．また，本タスクの評価基盤として，48件のテキスト分類タスクと20種類の事前学習済みBERTモデルからなる新たなデータセットを構築した．さらに，既存の主要なモデル検索手法をベースラインとして評価するとともに，Prompt Tuningによるタスク表現とモデルの埋め込みベクトルをTransformer Encoderで統合する新たなモデル検索手法を提案する．

なお，本研究で定義している事前学習済みBERTモデル検索タスクは，国際会議NTCIR-19 [1]において，*ModelRetrieval*タスクのサブタスクとして実施している [2]．

1：https://research.nii.ac.jp/ntcir/ntcir-19/index-ja.html
2：https://modeltreival.github.io/modelretrieval-1/

## 2 関連研究

### 2.1 事前学習済みモデルとドメイン特化型モデル

現代の機械学習において，下流タスクに適応させた事前学習済みモデルの利用は不可欠となっている．自然言語処理分野では，BERT [9] が Transformer アーキテクチャ [24] を用いた大規模事前学習の有効性を確立した．その後継モデルは，さらなる効率と精度の向上を実現している．例えば，RoBERTa は Next Sentence Prediction（NSP）タスクを排除し，データ規模を拡大した [17]．ALBERT は，埋め込み行列の分解と層間でのパラメータ共有によりモデルパラメータを削減した [14]．また，DeBERTa は Disentangled Attention と強化されたデコーディング機構を導入している [12]．さらに，DistilBERT [23] のように，知識蒸留を用いて性能を維持しつつモデルを軽量化する手法も提案されている．

ドメイン特化は，転移学習の効果をさらに高める．具体例として，科学文献向けの SciBERT [3]，金融ドメイン向けの FinBERT [18]，攻撃的な表現を扱う HateBERT [5]，そして法務コーパス向けの LEGAL-BERT [7] などが挙げられる．生物医学分野では，BioBERT [15]，ClinicalBERT [13]，MedBERT [21] などがドメイン固有のコーパスを活用している．また，インターネット上のテキスト分布に適応させたモデルとして IMHO [6] や BERTweet [19] がある．ドメイン適応事前学習（Domain-adaptive pre-training）は，ラベルなしデータを用いて一般コーパスから新規ドメインへの橋渡しを行う有効な手段である [11]．

### 2.2 機械学習モデル検索 (Machine Learning Model Retrieval)

新たなタスクに対して適切なモデルを選択する研究は，「モデル選択（Model selection）」，「転移性推定（Transferability estimation）」，「ニューラルネットワーク検索（Neural network retrieval）」など様々な名称で呼ばれており，多岐にわたるタスク，分野，モダリティにまたがっている．

例えば自然言語処理分野では，Safikhani ら [22] がテキスト分類のためのドメインを考慮した事前学習済みモデル検索を研究しており，Dai ら [8] は検索タスク（Retrieval tasks）における転移性を評価している．コンピュータビジョン分野では，Bolya ら [4] や Nermeen ら [1] が画像分類モデルの検索を対象としているほか，Fouquet ら [10] は物体検出，Yang [25] はセグメンテーションタスクを扱っている．

また，情報検索の国際会議である TREC 2025 においても，Million LLMs Track [3]が実施されている．同トラックは，ユーザの任意の質問（プロンプト）に対して最適な回答を生成する大規模言語モデル（LLM）を検索するタスクであり，生成 AI 時代におけるモデル選択の重要性が広く認識されつつあることがわかる．

先行研究ではタスクやモダリティによって異なる用語が用いられているが，これらは「与えられたタスクと制約条件の下で，候補モデルを期待される性能順にランク付けする」という共通の目的を持っている．本研究では，この視点を「機械学習モデル検索（Machine Learning Model Retrieval）」と定義する．

3：https://trec-mllm.github.io/

## 3 問題定義

事前学習済み BERT モデル検索タスクを定義する．まず，検索対象となる $K$ 個の事前学習済みモデルの集合を $\mathcal{M}$ とする．

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$$

ここで，各モデル $m_k$ は特定のアーキテクチャと事前学習済みのパラメータを持つ．

次に，検索クエリとなるターゲットタスク $\mathcal{T}$ を定義する．本研究において，タスク $\mathcal{T}$ は訓練データ，検証データ，テストデータの 3 つの集合から構成されるとする．

$$\mathcal{T} = \{D_{\text{train}}, D_{\text{val}}, D_{\text{test}}\}$$

ここで，各データセットは入力文書 $x$ とラベル $y$ のペアの集合 $\{(x_i, y_i)\}_{i=1}^{N}$ である．

モデル検索タスクの目的は，ターゲットタスク $\mathcal{T}$ に対して，最も高い性能を発揮するモデル $m \in \mathcal{M}$ を特定することである．

あるモデル $m_k$ を $\mathcal{T}$ の訓練データ $D_{\text{train}}$ および検証データ $D_{\text{val}}$ を用いてファインチューニングして得られるモデルを $m'_k$ とする．このとき，テストデータ $D_{\text{test}}$ におけるモデルの真の性能を $S(m'_k, D_{\text{test}})$ と定義する．本研究では評価指標として Macro F1 値を用いる．

検索システム $f$ は，テストデータ $D_{\text{test}}$ を参照することなく，以下の推定スコア $\hat{s}_k$ を算出するものと定義する．

$$\hat{s}_k = f(m_k, D_{\text{train}}, D_{\text{val}})$$

算出されたスコア $\hat{s}_k$ に基づくモデルのランク付けが，真の性能 $S(m'_k, D_{\text{test}})$ に基づくランク付けに近いほど検索性能が良い．

## 4 事前学習済み BERT モデル検索タスクのための評価データセット

本研究では，事前学習済み BERT モデル検索タスクの評価用データセットとして，48 件の文書分類タスクと 20 種類の異なる事前学習済み BERT モデルからなるベンチマークを構築した．各タスクに対して全モデルを適用し，正解となるランクリストを作成した．この正解リストと検索結果を比較することで，検索手法の定量的な評価が可能となる．

### 4.1 文書分類タスク

ベンチマークデータセットに含まれる文書分類タスクの一部を表 1 に示す．これらのタスクは Hugging Face Datasets で公開されており，`https://huggingface.co/datasets/<dataset_name>` からアクセス可能である．各タスクはラベル数やデータ規模が

異なっている．

各タスクのデータセットは，**訓練データ**，**検証データ**，**テストデータ**の 3 つに分割して使用する．各タスクのデータ分割は以下のルールに基づいて行った．

- Hugging Face 上ですでに訓練・検証・テストデータに分割されている場合：その分割をそのまま利用する．
- 訓練セットとテストセットのみに分割されている場合：元の訓練データの 10%を検証データとして切り出し，残りを訓練データとする．
- 訓練データのみが提供されている場合：全データの 10%を検証用，10%をテスト用として切り出し，残りの 80%を訓練データとする．

さらに，実験における計算コストを抑制するため，各データ分割のサイズに上限を設定した．具体的には，上記の手順で分割された訓練，検証，およびテストデータのサンプル数がそれぞれ 5,000 件を超える場合，ランダムサンプリングによって各 5,000 件にした．

### 4.2 検索対象 BERT モデル

検索対象となる事前学習済み BERT モデルの一部を表 2 に示す．これらのモデルも Hugging Face で公開されており，`https://huggingface.co/<model_name>` から利用可能である．各モデルは，アーキテクチャ（BERT，RoBERTa など），パラメータ数，および事前学習やファインチューニングに使用されたデータセットが異なっており，これらが特定の文書分類タスクに対する適合性に影響を与えると考えられる．例えば，ツイートデータで学習されたモデルは，一般的な Web テキストで訓練されたモデルと比較して，ツイート関連タスクにおいてより高い性能を発揮する可能性がある．

### 4.3 正解ランクリストの作成

各タスクに対する事前学習済み BERT モデルの正解ランクリスト $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ は，以下の手順で作成した．

1. タスクの訓練データと検証データを用いて，すべての事前学習済みモデルに対してファインチューニングを行う．
2. タスクのテストデータを用いて，ファインチューニング済みモデルの性能を Macro F1 値で評価する．
3. テストデータにおける Macro F1 値に基づいて，モデルを降順にランク付けする．

実験環境として，4 基の NVIDIA RTX 3090 GPU (24GB) を使用した．ハイパーパラメータは，バッチサイズを 16，学習率を $2 \times 10^{-5}$，最適化手法に AdamW，Weight Decay を 0.01 とした．損失関数にはクロスエントロピー誤差を用い，検証データの損失に基づいた Early Stopping（patience=10）を適用した．

### 4.4 タスクの重要度と分類

モデル選択の難易度や重要性はタスクごとに異なる．そこで，各タスクにおけるモデル検索の重要度を定量化するため，**期待リグレット（Expected Regret）**を導入する．

まず, タスク $\mathcal{T}$ におけるモデル $m$ の Macro F1 値を $S(m, \mathcal{T})$ とし, タスク内の最大性能で正規化した値を正規化性能 $\bar{S}(m, \mathcal{T})$ と定義する．

$$\bar{S}(m, \mathcal{T}) = \frac{S(m, \mathcal{T})}{\max_{m' \in \mathcal{M}} S(m', \mathcal{T})} \tag{1}$$

この正規化性能を用いて，候補モデル集合 $\mathcal{M}$ からランダムにモデルを選択した場合に，最適なモデルと比較して平均的にどの程度の性能損失が生じるかを表す期待リグレット $R_{\text{exp}}(\mathcal{T})$ を以下のように定義する．

$$R_{\text{exp}}(\mathcal{T}) = 1 - \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \bar{S}(m, \mathcal{T}) \tag{2}$$

$R_{\text{exp}}$ が大きいタスクは，モデルによる性能差が大きく，適切なモデルを選択できなかった場合の損失が大きいため，モデル検索の重要度（Criticalness）が高いと言える．本研究では，このスコアに基づいてタスクを以下の 3 つのカテゴリーに分類した．

- **Low criticalness**: $R_{\text{exp}} < 0.03$
- **Medium criticalness**: $0.03 \leqq R_{\text{exp}} < 0.10$
- **High criticalness**: $0.10 \leqq R_{\text{exp}}$

### 4.5 実験設定と評価指標

構築した 50 件のタスクを，Low, Medium, High criticalness の各カテゴリーの割合が均等になるように，検索モデル学習用の訓練タスクセット（25 件）と評価用のテストタスクセット（25 件）に分割した．

検索手法の評価指標として，nDCG@$k$（Normalized Discounted Cumulative Gain）を用いる．上位 $k$ 件の検索結果に対する DCG@$k$ は以下のように計算する．

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)} \tag{3}$$

ここで，順位 $i$ のモデルの適合スコア $\text{rel}_i$ は，そのモデルの正規化性能 $\bar{S}$ に基づいて以下のように定義する．

$$\text{rel}_i = \begin{cases} 4 & (0.975 \leqq \bar{S} \leqq 1.000) \\ 3 & (0.950 \leqq \bar{S} < 0.975) \\ 2 & (0.925 \leqq \bar{S} < 0.950) \\ 1 & (0.900 \leqq \bar{S} < 0.925) \\ 0 & (0.000 \leqq \bar{S} < 0.900) \end{cases} \tag{4}$$

最終的な nDCG@$k$ は，理想的なランキングの IDCG@$k$ を用いて計算する．

表 1 本研究で構築したベンチマークデータセットに含まれる一部の文書分類タスク．データ数は前処理（上限 5,000 件）後の最終的な事例数を示す．タスク名の右肩の記号はデータ分割方法の差異を表す（無印：公式分割，*：訓練データから検証データを分割，**：訓練データから検証・テストデータを分割）．

| ID | タスク名 | サブセット | ラベル数 | データ数 | | |
|---|---|---|---|---|---|---|
| | | | | 訓練 | 検証 | テスト |
| 1 | cardiffnlp/tweet_eval | emoji | 20 | 5,000 | 5,000 | 5,000 |
| 17 | stanfordnlp/sst2** | – | 2 | 5,000 | 5,000 | 5,000 |
| 50 | toxigen/toxigen-data* | annotated | 5 | 5,000 | 896 | 940 |
| ... | ... | ... | ... | ... | ... | ... |

表 2 検索対象となる事前学習済み BERT モデルの一部．各モデルのアーキテクチャおよび事前学習に使用されたデータセット情報を示す．

| ID | モデル名 | 事前学習データ | 事前ファインチューニング |
|---|---|---|---|
| | **BERT 系モデル** | | |
| 4 | GroNLP/hateBERT | BookCorpus, English Wikipedia, RAL-E | – |
| ... | ... | ... | ... |
| | **DistilBERT 系モデル** | | |
| 11 | distilbert-base-uncased-distilled-squad | BookCorpus, English Wikipedia | SQuAD |
| ... | ... | ... | ... |
| | **RoBERTa 系モデル** | | |
| 15 | zhayunduo/roberta-base-stocktwits-finetuned | BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories | StockTwits |
| ... | ... | ... | ... |
| | **DistilRoBERTa 系モデル** | | |
| 18 | j-hartmann/emotion-english-distilroberta-base | BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories, Tweets | Crowdflower, Emotion, GoEmotions, ... |
| ... | ... | ... | ... |
| | **その他のモデル (ALBERT, DeBERTa)** | | |
| 20 | microsoft/deberta-base | BookCorpus, English Wikipedia, OpenWebText, Stories | – |
| ... | ... | ... | ... |

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k} \tag{5}$$

本研究では，$k \in \{1, 3\}$ で評価を行う．

## 5 検 索 手 法

本節では，事前学習済みモデルの検索手法を分類し，実験に用いる既存手法および提案手法ついて説明する．

### 5.1 検索手法の分類

既存のモデル検索手法は，検索時に候補モデルへのデータ入力（推論）を必要とするか否かによって，主に**推論あり（Inference-based）**と**推論なし（Inference-free）**の 2 種類に分類することができる．

a） 推論ありの手法

クエリタスクのデータ $\mathcal{T}$（訓練データ $D_{\text{train}}$ 等）を候補モデル $m \in \mathcal{M}$ に入力し，得られた特徴量 $f(x; m)$ を用いてスコアを算出する手法である．候補モデル数 $K$ に比例して計算コストが増大するが，タスクとモデルの適合性をデータに基づいて直接的に評価できるため，推定精度が高い傾向にある．

b） 推論なしの手法

検索時に，クエリタスクのデータを候補モデルに入力することなく検索を行う手法である．モデルの特性を表すベクトルを事前に計算しておきタスク特性との類似度を計算するアプローチや，過去のタスクにおける性能統計のみを利用するアプローチなどがある．個々のモデルに対する推論コストが発生しないため，高速な検索が可能である．

### 5.2 既存手法手法

#### 5.2.1 LogME

LogME (Logarithm of Maximum Evidence) [26] は，**推論あり**の代表的な手法である．本手法は，事前学習済みモデルを固定的な特徴抽出器と見なし，その出力特徴量に対する線形回

帰モデルの周辺尤度（Evidence）をスコアとして用いる．

具体的には，モデル $m$ を用いて訓練データ $D_{\text{train}}$ および検証データ $D_{\text{val}}$ から抽出した特徴行列を $\mathbf{F}$，ラベルに対応するターゲット（分類タスクの場合は One-hot ベクトル等）を $\mathbf{y}$ とする．線形回帰の重みを $\mathbf{w}$，観測ノイズの精度を $\beta$，重みの事前分布の精度を $\alpha$ としたとき，LogME スコア $S_{\text{LogME}}(m)$ は，以下の周辺尤度 $p(\mathbf{y}|\mathbf{F},\alpha,\beta)$ の最大値として定義される．

$$S_{\text{LogME}}(m) = \max_{\alpha,\beta} \log \int p(\mathbf{y}|\mathbf{F},\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w} \tag{6}$$

この値が大きいほど，モデルの特徴空間がタスクのラベル予測に適しており，ファイチューニング後の性能が高くなるとされる．

**5.2.2** H-Score

H-Score [2] は，特徴空間における情報の分離度合いを測定する**推論あり**の手法である．この手法は，特徴量の冗長性を考慮しつつ，クラス間の分散が全分散に対してどれだけの割合を占めるかを評価する．

モデル $m$ によって抽出されたデータ $x_i$ の特徴量を $\mathbf{f}_i = f(x_i;m)$ とする．全データの平均ベクトルを $\boldsymbol{\mu}$，全共分散行列を $\boldsymbol{\Sigma}_{\text{tot}}$ とする．また，クラス $c$ に属するデータの平均ベクトルを $\boldsymbol{\mu}_c$，$N_c$ をクラス $c$ のデータ数としたとき，クラス間共分散行列 $\boldsymbol{\Sigma}_B$ は以下のように定義される．

$$\begin{aligned}\boldsymbol{\Sigma}_{\text{tot}} &= \frac{1}{N}\sum_i (\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^\top \\ \boldsymbol{\Sigma}_B &= \sum_c \frac{N_c}{N}(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top\end{aligned} \tag{7}$$

H-Score $S_{\text{H}}(m)$ は，全共分散行列の逆行列とクラス間共分散行列の積のトレースとして定義される．

$$S_{\text{H}}(m) = \text{Tr}(\boldsymbol{\Sigma}_{\text{tot}}^{-1}\boldsymbol{\Sigma}_B) \tag{8}$$

値がほどタスクに適していることを示す．

**5.2.3** k 近傍法

k 近傍法（k-NN）を用いた手法 [20] は，モデルの特徴空間上での近傍探索により性能を推定する**推論あり**の手法である．

まず, モデル $m$ を用いて, 訓練データ $(x_i^{\text{train}}, y_i^{\text{train}}) \in D_{\text{train}}$ および検証データ $(x_j^{\text{val}}, y_j^{\text{val}}) \in D_{\text{val}}$ の特徴ベクトル $\mathbf{d}_i^{\text{train}}, \mathbf{d}_j^{\text{val}}$ をそれぞれ計算する．次に，訓練データの特徴ベクトルを参照集合として，検証データの各事例に対して k-NN による分類を行う．

検証事例 $j$ に対する予測ラベルを $\hat{y}_j$ としたとき，スコア $S_{\text{kNN}}(m)$ は検証データセット全体に対する予測の Macro F1 値として定義される．

$$S_{\text{kNN}}(m) = \frac{1}{|D_{\text{val}}|}\sum_{j=1}^{|D_{\text{val}}|} \mathbb{I}(\hat{y}_j = y_j^{\text{val}}) \tag{9}$$

ここで，$\mathbb{I}(\cdot)$ は条件が真の場合に 1，偽の場合に 0 を返す指示関数である．

**5.2.4** ModelSpider

ModelSpider [27] は，検索時に対象モデルへの推論を必要としない**推論なし**の手法である．本手法では，各事前学習済みモデル $m \in \mathcal{M}$ の特性を表すモデルベクトル $\mathbf{v}_m$ を事前に抽出・保存しておく．

検索時には，まずクエリタスクの訓練データ $D_{\text{train}}$ からタスクの特性を表す**タスク行列** $\mathbf{V}_{\text{task}}$ を構築する．これは，各クラスラベル $c$ に属するサンプルの特徴量の平均ベクトル（重心）$\mathbf{c}_c \in \mathbb{R}^d$ を行ベクトルとして積み重ねた行列として定義される．

$$\mathbf{V}_{\text{task}} = [\mathbf{c}_1, \ldots, \mathbf{c}_C]^\top \in \mathbb{R}^{C\times d} \tag{10}$$

ここで $C$ はクラス数，$d$ は特徴量の次元数である．

タスクに対するモデルのスコア $S_{\text{Spider}}(m)$ は，タスク行列 $\mathbf{V}_{\text{task}}$ とモデルベクトル $\mathbf{v}_m$ を入力とし，学習済みのスコアリング関数 $\phi$ を用いて以下のように算出される．

$$S_{\text{Spider}}(m) = \phi(\mathbf{v}_m, \mathbf{V}_{\text{task}}) \tag{11}$$

ここで関数 $\phi$ は，1 層の Transformer Encoder および全結合層からなるニューラルネットワークである．

**5.2.5** Average Rank

Average Rank は，訓練タスクにおけるモデルの性能統計のみを利用する**推論なし**のベースライン手法である．具体的には，訓練タスク $\mathcal{T}_{\text{train}}$ を用いて，各モデル $m \in \mathcal{M}$ の平均ランクをスコア $S_{\text{Avg}}(m)$ として算出する．

$$S_{\text{Avg}}(m) = \frac{1}{|\mathcal{T}_{\text{train}}|}\sum_{\mathcal{T}\in\mathcal{T}_{\text{train}}} \text{Rank}(m,\mathcal{T}) \tag{12}$$

ここで $\text{Rank}(m,\mathcal{T})$ はタスク $\mathcal{T}$ におけるモデル $m$ の正解順位である．検索時には，ターゲットタスクの内容（テキストやラベル）に依存せず，常にこの $S_{\text{Avg}}(m)$ の昇順（平均順位が良い順）にソートされた固定のランクリストを出力する．

### 5.3 提案手法

本研究では，新たな**推論なし**の手法を提案する．図 1 に示すように，本手法では，クエリとなるターゲットタスクの特徴量と，検索候補となる各モデルの埋め込みベクトルを統合し，その適合度スコアを推定するものである．以下，学習フェーズと推論フェーズについて詳細を述べる．

**5.3.1** 学習フェーズ

学習フェーズでは，タスクとモデルのペアを入力として，そのタスクに対するモデルの性能を精度よく予測するように最適化を行う．

a） タスクエンベディング

クエリタスク $\mathcal{T}$ の特徴量を抽出するため，Zhou らによって提案された TuPaTE [28] に基づく手法を用いる．具体的には，

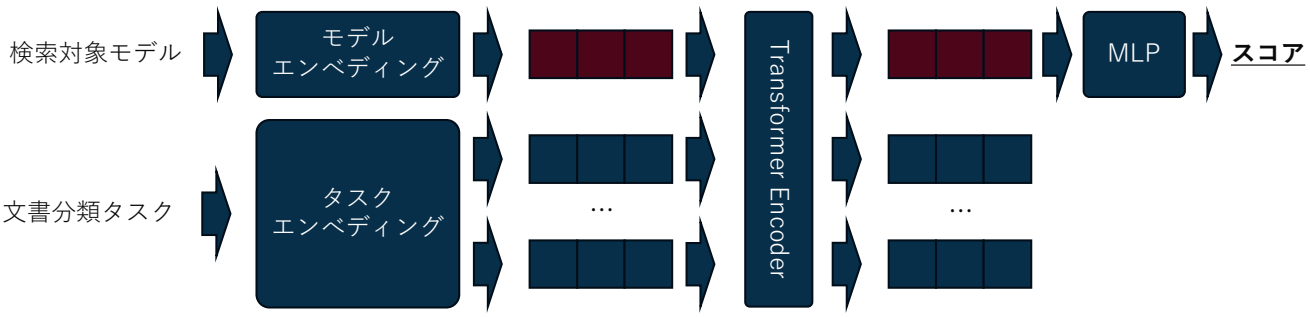



図 1　提案手法の概要：クエリタスクから Prompt Tuning により抽出されたタスクエンベディングと，各候補モデルの学習可能な埋め込みベクトルを Transformer Encoder で統合し，MLP を介して適合度スコアを予測する．

任意の言語モデルに対し，Prompt Tuning [16] を適用する．

まず，タスクの入力テキストの先頭に $P = 20$ 個の学習可能な特殊トークン（Soft Prompts）を付加し，拡張された入力系列を構成する．この系列をモデルに入力し，得られた `[CLS]` トークンのベクトルを分類器に通してラベルを予測する．この際，Transformer のパラメータは固定し，特殊トークンの埋め込み行列と分類器のみを更新対象とする．学習後，得られた $P$ 個の特殊トークンの埋め込みベクトル集合を，タスク $\mathcal{T}$ の特徴量行列 $\mathbf{V}_{\mathcal{T}}^{(\mathrm{tsk})} \in \mathbb{R}^{P\times 768}$ とする．

本研究では, 事前学習済みの `bert-base-multilingual-uncased` に対しクエリタスクの訓練データを用いて Prompt Tuning を行った．エポック数を 20、最適化を AdamW、学習率を 1e-3 とした．Prompt Tuning の学習時間は一つのタスクにおいて平均として約 2 分間程度であった．

b） モデルエンベディング

検索対象となる各モデル $m_k \in \mathcal{M}$ に対して，ランダムに初期化された学習可能な埋め込みベクトル $\mathbf{v}_k^{(\mathrm{mdl})} \in \mathbb{R}^{768}$ を割り当てる．

c） スコア予測

あるタスク $\mathcal{T}$ とモデル $m_k$ の適合度を算出するため，タスクエンベディングとモデルエンベディングを結合した入力を構成する：

$$\mathbf{Z}_{k,\mathcal{T}} = \left[\mathbf{v}_k^{(\mathrm{mdl})}; \mathbf{V}_{\mathcal{T}}^{(\mathrm{tsk})}\right] \in \mathbb{R}^{(P+1)\times 768} \tag{13}$$

この入力に対し，位置エンコーディングを付加した後，2 層の Transformer Encoder に入力する．Encoder の出力のうち，モデルベクトルに対応する位置のベクトルを $\mathbf{u}_{k,\mathcal{T}} \in \mathbb{R}^{768}$ とする．最後に，$\mathbf{u}_{k,\mathcal{T}}$ を MLP に入力し，推定スコア $\hat{s}_k$ を算出する：

$$\hat{s}_k = \mathrm{MLP}(\mathbf{u}_{k,\mathcal{T}}) \tag{14}$$

d） 最　適　化

予測スコア $\hat{s}_k$ と，実際のファインチューニングによって得られた真の Macro F1 値 $s_k$ との間の平均二乗誤差（MSE）を損失関数として定義する：

$$L = \frac{1}{|\mathcal{M}|}\sum_{k=1}^{K}(\hat{s}_k - s_k)^2 \tag{15}$$

この損失を最小化するように，モデルエンベディング $\{\mathbf{v}_k^{(\mathrm{mdl})}\}$，Transformer Encoder，および MLP の各パラメータを誤差逆伝播法により更新する．

### 5.4　推論フェーズ

推論フェーズでは，未知のターゲットタスクに対し，以下の手順でモデルのランキングを生成する．

1. **タスクエンベディングの抽出**: 新たなタスク $\mathcal{T}$ の訓練データを用い，Prompt Tuning によって $P$ 個の特殊トークンベクトルを取得する．
2. **スコア推定**: 学習済みの各モデルエンベディング $\mathbf{v}_k^{(\mathrm{mdl})}$ とタスクエンベディングを順次結合し，学習済みの Transformer Encoder および MLP を通じて各モデルの予測スコア $\hat{s}_k$ を算出する．
3. **ランキングの生成**: 算出された $\hat{s}_k$ の降順にモデルをソートし，ターゲットタスクに適合する可能性が高い順にモデルリストを提示する．

## 6　評 価 結 果

本節では，構築したデータセットにおける提案手法および既存の検索手法（KNN，LogME，H-Score，Model Spider，Avg. Rank）の検索性能を評価する．表 3 に評価結果の詳細を示す．

検索重要度（Criticalness）が High なタスク群においては，各手法間で性能の差が顕著に見られた．全体的な傾向として，推論ありの手法の方が，推論なしに比べて高い検索精度を示す傾向にある．一方で，推論ありの手法は検索時の計算コストが高いことから，実用上は検索速度との間にトレードオフが存在する点に留意が必要である．

推論なしの手法群に着目すると，検索重要度が高いタスクにおいて，提案手法が NDCG@1 (0.750) および NDCG@3 (0.719) の双方で最も高い平均スコアを記録した．今後，学習データがさらに増加する環境においては，提案手法を含む推論なし手法のさらなる性能向上が期待できる．

表 3　全タスクにおける各検索手法の定量評価結果（NDCG スコア）．タスク重要度（Criticalness）ごとの平均値および標準偏差を併記する．各項目における最高値を**太字**，次点を下線で示す．

| Group | ID | NDCG@1 | | | | | | NDCG@3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 推論なし | | | 推論あり | | | 推論なし | | | 推論あり | | |
| | | 提案手法 | M.Sp | AvR | KNN | LogME | H-Sc | 提案手法 | M.Sp | AvR | KNN | LogME | H-Sc |
| High | 3 | **1.000** | 0.000 | 0.000 | **1.000** | 0.500 | **1.000** | **0.765** | 0.000 | 0.235 | **0.685** | 0.656 | 0.564 |
| | 6 | **1.000** | 0.500 | 0.500 | **1.000** | **1.000** | **1.000** | **0.725** | 0.435 | 0.672 | 0.712 | 0.541 | **0.779** |
| | 9 | **0.750** | 0.500 | 0.500 | **1.000** | **1.000** | **1.000** | **0.712** | 0.491 | 0.644 | **1.000** | 0.899 | 0.920 |
| | 13 | 0.500 | **1.000** | **1.000** | **1.000** | 0.750 | 0.500 | 0.626 | 0.865 | **0.932** | **0.921** | 0.875 | 0.672 |
| | 28 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.852 | 0.676 | **0.926** | **0.883** | **0.883** | **0.883** |
| | 37 | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.750** | **0.824** | 0.352 | **0.824** | **0.809** | 0.528 | 0.528 |
| | 48 | 0.250 | **1.000** | **1.000** | 0.000 | 0.000 | **0.750** | 0.531 | 0.499 | **0.561** | 0.457 | 0.398 | **0.824** |
| | Mean | **0.750** | 0.679 | 0.679 | 0.821 | 0.679 | **0.893** | **0.719** | 0.474 | 0.685 | **0.781** | **0.781** | 0.641 |
| | STD | 0.289 | 0.374 | 0.374 | 0.374 | 0.134 | 0.374 | 0.112 | 0.269 | 0.244 | 0.182 | 0.142 | 0.189 |
| Medium | 10 | **1.000** | **1.000** | **1.000** | **0.750** | **0.750** | **0.750** | 0.926 | 0.661 | **1.000** | 0.809 | **0.824** | **0.824** |
| | 14 | **0.500** | 0.250 | 0.250 | **1.000** | 0.750 | **1.000** | **0.515** | 0.500 | 0.457 | **0.867** | 0.778 | 0.633 |
| | 17 | 0.750 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.883 | 0.735 | **0.941** | **0.883** | 0.765 | 0.765 |
| | 24 | **0.750** | 0.500 | 0.500 | **1.000** | 0.000 | **1.000** | 0.691 | 0.574 | **0.707** | **1.000** | 0.531 | **1.000** |
| | 43 | **1.000** | **1.000** | **1.000** | **1.000** | 0.750 | **1.000** | **1.000** | 0.941 | **1.000** | **1.000** | 0.809 | 0.867 |
| | 47 | **1.000** | 0.750 | 0.750 | **1.000** | **1.000** | 0.750 | 0.645 | 0.661 | **0.883** | **0.941** | 0.587 | 0.926 |
| | Mean | **0.833** | 0.750 | 0.750 | **0.958** | 0.667 | **0.958** | 0.777 | 0.679 | **0.831** | **0.917** | 0.860 | 0.691 |
| | STD | 0.204 | 0.316 | 0.316 | 0.102 | 0.342 | 0.102 | 0.188 | 0.152 | 0.213 | 0.077 | 0.091 | 0.124 |
| Low | 11 | **1.000** | **1.000** | **1.000** | **1.000** | 0.750 | **1.000** | **1.000** | 0.926 | 0.926 | **1.000** | 0.883 | **1.000** |
| | 16 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 20 | **1.000** | **1.000** | **1.000** | 0.500 | **0.750** | **0.750** | **1.000** | 0.867 | **1.000** | 0.691 | **0.735** | **0.735** |
| | 26 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 29 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.867 | **1.000** | **1.000** | **1.000** | **1.000** |
| | 32 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 34 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 35 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 39 | 0.750 | **1.000** | **1.000** | **0.750** | **0.750** | **0.750** | 0.883 | **1.000** | 0.941 | **0.883** | **0.883** | **0.883** |
| | 44 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** |
| | 45 | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.941 | **1.000** | **0.941** | **0.941** | **0.941** |
| | 46 | 0.750 | **1.000** | **1.000** | 0.500 | 0.500 | **1.000** | 0.707 | **1.000** | 0.941 | 0.765 | 0.765 | **0.883** |
| | Mean | 0.958 | **1.000** | **1.000** | **0.896** | **0.896** | **0.958** | 0.966 | 0.967 | **0.984** | 0.940 | 0.934 | **0.953** |
| | STD | 0.097 | 0.000 | 0.000 | 0.198 | 0.097 | 0.167 | 0.088 | 0.053 | 0.029 | 0.106 | 0.083 | 0.097 |

## 7 結　　論

本研究では，事前学習済み BERT モデル検索という問題を体系的に定義し，その評価基盤として 48 件の文書分類タスクと 20 種類のモデルからなるベンチマークデータセットを構築した．本データセットを用い，既存の 5 つの主要な検索手法に加え，タスク表現とモデル埋め込みをを用いた新たなモデル検索手法を提案し，比較評価を行った．

実験の結果，検索の重要度が高い **High Criticalness** なタスクにおいては，推論を伴う手法（LogME および k 近傍法）が高い検索精度を示す一方，推論を必要としない **Inference-free** な手法の中では，提案手法が最も優れた性能を達成することを確認した．これにより，検索時の計算コストを抑えつつ高精度なモデル選定を実現する手法としての有効性が示された．

今後の展望としては，まず検索対象となるモデルのバリエーションをさらに拡大し，より大規模なモデル群への対応を進める．また，クエリとなるタスクの拡張として，LLM を用いたデータ生成手法などを活用し，評価タスクのさらなる拡充を目指す．本研究で構築した評価基盤が，今後のモデル検索研究のさらなる発展に貢献することを期待する．

## 謝　　辞

本研究は，JSPS 科研費 JP24K03228，JP25K03229，JP25K03228，ならびに，2025 年度国立情報学研究所公募型共同研究（251S4-22794）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] Nermeen Abou Baker and Uwe Handmann. One size does not fit all in evaluating model selection scores for image classification. *Scientific Reports*, Vol. 14, , 2024.

[2] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An Information-Theoretic Approach to Transferability in Task Transfer Learning. In *Proceedings of 2019 IEEE International Conference on Image Processing (ICIP 2019)*, pp. 2309–2313, 2019.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pp. 3615–3620, 2019.

[4] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable Diverse Model Selection for Accessible Transfer Learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021*, pp. 19301–19312, 2021.

[5] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 17–25, 2021.

[6] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO Fine-Tuning Improves Claim Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*, pp. 558–563, 2019.

[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 2898–2904, 2020.

[8] Mengyu Dai, Amir Hossein Raffiee, Aashish Jain, and Joshua Correa. Evaluating Transferability in Retrieval Tasks: An Approach Using MMD and Kernel Methods. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 22390–22400, 2024.

[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 (NAACL 2019)*, pp. 4171–4186, 2019.

[10] Louis Fouquet, Simona Maggio, and Léo Dreyfus-Schmidt. Transferability Metrics for Object Detection. *arXiv preprint arXiv:2306.15306*, 2023.

[11] Xiaochuang Han and Jacob Eisenstein. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pp. 4238–4248, November 2019.

[12] He, Pengcheng and Liu, Xiaodong and Gao, Jianfeng and Chen, Wei. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, pp. 1–21, 2021.

[13] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[14] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*, 2019.

[15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Vol. 36, pp. 1234–1240, 2019.

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pp. 3045–3059, November 2021.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

[18] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. FinBERT: a pre-trained financial language representation model for financial text mining. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pp. 4513–4519, 2021.

[19] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 9–14, 2020.

[20] Huu-Long Pham, Ryota Mibayashi, Takehiro Yamamoto, Makoto P. Kato, Yusuke Yamamoto, Yoshiyuki Shoji, and Hiroaki Ohshima. Inference-based no-learning approach on pre-trained BERT model retrieval. In *Proceddings of the 2024 IEEE International Conference on Big Data and Smart Computing (BigComp 2024)*, pp. 234–241, 2024.

[21] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, Vol. 4, , 2021.

[22] Parisa Safikhani and David Broneske. AutoML Meets Hugging Face: Domain-Aware Pretrained Model Selection for Text Classification. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)*, pp. 466–473, 2025.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 6000–6010, 2017.

[25] Yuncheng Yang, Meng Wei, Junjun He, Jie Yang, Jin Ye, and Yun Gu. Pick the Best Pre-trained Model: Towards Transferability Estimation for Medical Image Segmentation. In *Proceedings of the 26th Medical Image Computing and Computer Assisted Intervention (MICCAI 202)*, pp. 674–683, 2023.

[26] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. In *Proceedings of the 38th International Conference on Machine Learning (PMLR 2021)*, Vol. 139, pp. 12133–12143, 2021.

[27] Yi-Kai Zhang, Ting-Ji Huang, Yao-Xiang Ding, De-Chuan Zhan, and Han-Jia Ye. Model Spider: Learning to Rank Pre-trained Models Efficiently. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS 2023)*, pp. 13692–13719, 2023.

[28] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Ef-

ficiently Tuned Parameters Are Task Embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 5007–5014, December 2022.