数値条件クエリに対する密検索モデルの性能評価

藤巻晴葵 [1]　加藤誠 [2]

本研究では，数値条件を含むクエリ（例：「価格が 500〜1000 ドルのノートパソコン」「研究開発費が 10 億円以上の企業」）に対する密検索モデルの性能を評価した．実験では，密検索モデルが数値条件を部分的に考慮できる一方で，理想的な性能には大きく及ばないことが明らかになった．また，数値条件の種類や表現形式によって検索性能にモデル間で顕著な差異が生じることを示した．さらに，事前学習中に獲得された知識が数値条件を含むクエリへの応答に影響を与え，不適切なバイアスを引き起こす可能性があることが確認された．本研究は，特に e コマースや医療，金融といった数値情報に関する情報要求が多い分野において，密検索モデルのさらなる性能向上が必要であることを示唆している．

1　はじめに

近年，密検索モデルが商品検索や金融文書の検索，医療文書の検索などの幅広い分野で利用されている．密検索モデルはクエリや文書を BERT [4] などの言語モデルを利用し，密ベクトルに変換，そのベクトル同士の類似度を元に検索するモデルであり，BM25 や TF-IDF といった従来の検索手法に比べ，クエリや文書に含まれる文脈を考慮した検索が可能である．そのため，商品検索や金融・医療などの文書検索に用いられる．この文脈を考慮した検索を可能にするのは密検索モデルで利用している言語モデルが与えられた文章の単語同士の関係性を考慮して処理しているためである．一方，言語モデルは数値の理解が不十分であることが以前から指摘されている [5, 17]．このことから密検索モデルは数値を含んだクエリの検索に対して十分な性能を出すことができないことが考えられる．

数値を含んだクエリの例としては，「iPhone 16 のカメラ性能」というような数値クエリや「縦の長さが 10cm 以上のスマートフォン」，「胃のポリープの大きさが 5mm から 10mm だった患者のカルテ」というような数値条件を含むクエリが存在する．数値を含んだクエリは商品検索や金融文書，医療文書の検索でも行われる可能性があり，密検索モデルはそのクエリに対して適合する文書を正しく返すことができないことが懸念される．

本研究では，数値を含んだクエリの中でも数値条件クエリに焦点を当て，密検索モデルの数値条件クエリに対する性能に対して調査する．また，今後は数値条件クエリを **NumQ** と省略して示す．本研究では，上記の調査において 3 つのリサーチクエスチョン（RQ）を設定した．

- **RQ1**: 密検索モデルは **NumQ** に対して有効に働くか？
- **RQ2**: どのような種類の **NumQ** に対して有効に働くか？
- **RQ3**: 言語モデルの内部知識は **NumQ** の検索に影響を及ぼすのか？

RQ1 では，密検索モデルが数値条件を含む検索にどの程度有効であるのかを DPR [9]，E5 [18]，ColBERT [10, 15] などの主要な密検索モデルに対して，実データと合成データを組み合わせて作成した複数のジャンルのデータセットを利用して検証を行った．検証方法としては，数値条件を含むクエリとそうでないクエリで検索し，どの程度密検索モデルが数値条件を含むクエリの検索に有効であるのかを検証した．また，数値条件を含むクエリの検索結果から数値条件に適したもののみにフィルターした結果とも比較を行い，理想的な数値条件検索の結果とどの程度差があるのかについても明らかにした．実験結果から，密検索モデルはある程度数値条件を考慮した検索を行えることが示された．一方，理想的な性能とは大きな乖離があり，改善の余地が多く残されていることが明らかになった．

RQ2 では，どのような数値条件クエリに対して有効または有効ではないのかを 3 つの側面から検証した：数値条件の種類（Equal, Less than, More than, Around, Between）の違いによって検索有効性に違いがあるのか．時間や重量，長さなどの数値単位の違いによって検索有効性に変化はあるのか．単位，スケールなどの数値の表現方法（「5000 dollars」，「$5000」，「5k dollars」など）の違いによって，検索有効性がどのように変化するのか．実験結果から，数値条件や数値部分の表現，数値の単位の違いによって異なるモデル間で検索の有効性は変化することも明らかになった．

RQ3 では，密検索の学習に使用される言語モデルの学習時点から得られる知識が数値条件クエリの検索においてどの程度影響を及ぼすのかを 4 つの側面から調査を行った：映画のタイトルなどの固有名詞がそれに付随する数値条件クエリに対してどの程度影響があるのか．一般常識が **NumQ** に対して影響があるのか．国籍や性別などのセンシティブな属性が資格スコアや年収に関する **NumQ** にどの程度影響があるのか．密検索モデルの基盤モデルの違いによって，数値条件クエリの検索性能にどのような変化があるのか．実験結果から，言語モデルの事前学習で獲得された固有名詞に対する知識や一般常識などが **NumQ** の検索結果に影響を及ぼし，偏った検索結果につながる可能性があることを示唆していることが分かった．また，基盤モデルを数値に関連したデータセットや数値に特化した学習手法によって学習された言語モデルに変更し，密検索モデルを学習させて検索性能の変化を調査し，基盤モデルを変更することによって検索性能に変化があることが分かった．そして，各密検索モデルにおける性能の優劣はデータセットによって変化し，いずれの検索性能も理想的な性能とは程遠く，更なる改善が必要であることが示唆された．

本研究の貢献は以下の 4 点である．

- 数値条件を含む検索クエリに対する密検索モデルの有効性を複数のデータセットで評価した．
- 数値条件の種類や数値表現の違いが密検索モデルの性能に与

[1] 学生会員　筑波大学情報学位プログラム
fujimaki.haruki.tkb_eg@u.tsukuba.ac.jp
[2] 正会員　筑波大学図書館情報メディア系
mpkato@acm.org

える影響を明らかにした．
- 固有名詞や一般常識といった要因が数値条件を含むクエリの検索性能に及ぼす影響を明らかにした．
- 表データと文章テンプレートを対応付け，実データとテンプレート合成データを統合したベンチマークを構築し，数値条件に関連する広範な要因の評価を可能にした．

本論文の構成は以下の通りである．2 節では数値条件を含む検索に関する関連研究，および，密検索モデルとそのモデルに対する既知の問題について述べる．3 節では調査を行うための実験設定と各 RQ に関する実験結果について述べる．最後に，4 節では今後の課題と共に本論文の結論を述べる．

2 関連研究

本節では，数値条件クエリに対する検索システムと密検索モデルに関する関連研究について述べ，既存研究の位置づけを明確にする．

2.1 言語モデル

言語モデルは密検索モデルの基盤モデルとして利用されており，言語モデルの性能は密検索モデルの性能にも大きく関与している．言語モデルとして代表されるのが BERT [4] である．BERT は，Transformer [16] アーキテクチャを基盤とし，2 つの教師なし学習手法によって学習される．BERT は自然言語処理において高い性能を示す一方で，数値の理解に関する研究では，数値の理解が不十分であることが指摘されている．

Wallace ら [17] の研究では，自然言語処理モデルの数値理解能力に関する分析を行っている．数値の大きさや順序を正確に捉えられるかを検証しており，BERT などを含む複数のモデルを対象に数値デコーディング，加算タスクについて調査し，その結果，学習範囲外での数値に対して性能が大幅に低下することを示した．Dua ら [5] は，DROP (Discrete Reasoning Over Paragraphs) データセットを導入し，言語モデルの数値的推論能力を評価する新たなベンチマークを提案した．このベンチマークでは，テキスト中の数値に対する加算，カウント，ソートなどの離散的推論能力を測定し，既存のモデルの性能が低いことを示した．特に，人間の専門家と比較して大きな乖離があることが示されており，言語モデルにおける数値推論の課題を明らかにした．これらの研究は言語モデルの数値に関連した能力について検証を行なっているが，数値条件に対しては十分に検証されていない．

加えて，言語モデルが数値に対して理解が不十分であることを克服させるため，いくつかの改善手法が提案されている．FinBERT [2] は，金融ドメイン専用に設計された BERT ベースの言語モデルで，金融ドメインのコーパスで追加学習することで，金融特有の専門用語や文脈をより深く理解できるモデルとなっている．金融に関するベンチマークでは従来のモデルを大きく上回る性能を示している．SEC-BERT [13] は，財務文書に特化した BERT ベースの言語モデルで，金融に関するタスクに利用されることを目的としている．SEC-BERT-BASE は標準的な BERT-BASE アーキテクチャを持ち，財務文書で学習されたモデルである．また，数値トークンの処理方法に工夫を凝らした SEC-BERT-NUM や SEC-BERT-SHAPE などのバリエーションも存在する．GenBERT [6] は，数値的推論スキルを言語モデルに付与することを目的とした研究から作成されたモデルである．トークナイズ手法の変更や合成データを用いたマルチタスク学習により，DROP データセットでの F1 スコアを大幅に改善した．また，数値的推論能力を向上させながら，言語理解タスクでの性能も維持できることが特徴である．

これらの研究は，言語モデルの数値理解能力の向上に貢献しているが，数値条件クエリ（**NumQ**）に対する文書検索において有効に働くのかという観点では，まだ検証がされていない．また，既存研究では，数値のみの理解に焦点が当てられたものも多く，数値条件の複雑さや多様性を考慮した検証が十分ではない．本研究では，これらの課題を踏まえ，密検索モデルにおける **NumQ** の有効性を調査する．

2.2 数量中心の検索システム

数量中心の検索システムは，事前に文書から数値要素を抽出し，数値条件を数値条件クエリ（**NumQ**）から読み取り，照合することによって **NumQ** を処理する．Ho ら [7] は数量，単位，エンティティを含むタプルをクエリとコーパスから抽出し，そのタプルの類似度に基づいて数値条件の検索を行っている．加えて，同様のアプローチでウェブテーブルからタプルを抽出しクエリとの検索を行うシステムを提案している [8]．これらの研究では，数値情報に特化した追加のパイプラインを導入することによって **NumQ** の処理を可能にしているが，本研究では既存の密検索モデルを利用した検索システムの **NumQ** に対する能力に注目している．近年では，Almasian ら [1] が密検索モデルなどを含むニューラル検索モデルに対して，数値・単位に関する追加学習と数量抽出器を用いたリランキング手法による数量中心の検索システムを提案している．

本研究は密検索モデルを利用した検索システムの **NumQ** に対する検索有効性を調査するという点で部分的に重複しているが，表データとテンプレートから生成された実データと合成データによる評価により数値条件に関連した要因を広範かつ体系的に分析した．具体的には，以下の 3 点で先行研究を拡張する．（1）異なるタイプの **NumQ** に対する密検索モデルの検索有効性の調査．（2）言語モデルが **NumQ** に対する検索結果に与える影響の調査．（3）数値情報に特化した言語モデルが，**NumQ** に対する密検索モデルの検索有効性を改善することができるのか．

3 実験

本節では，設定した RQ に基づき，**NumQ** に対する密検索モデルの性能について調査するための方法の説明とその結果を述べる．

3.1 実験設定

NumQ の密検索モデルの性能を調査するため，実データと合成データの両方を用いてデータセットを構築した．各データセットの概要は表 1に示す．

使用した実データには，映画のタイトル，ジャンル，制作会社，興行収入を含む Movie dataset [3]，および LinkedIn から収集された求人情報と企業の従業員数を含む LinkedIn dataset [11] が含ま

表1　各データセットの概要

	Movie Rev.	Job Post	Comp. Emp.	Unit	Corp. Value	TOEFL Score	Job Salary
Research questions	1, 3	1, 3	1, 3	2	2	1, 3	3
Source	Movie	LinkedIn	LinkedIn	Synthetic	Synthetic	Synthetic	Synthetic
# of documents	6,048	17,614	16,287	134,919	53,946	110,000	110,000
# of queries	30,240	25,000	25,000	134,865	270,000	25,000	25,000

表2　Movie Revenue データセットにおける各数値条件のクエリテンプレート例

Equal	What are the movies that earned equal to {Number} in revenue?
Less	What are the movies that earned less than {Number} in revenue?
More	What are the movies that earned more than {Number} in revenue?
Around	What are the movies that earned revenue around {Number} in revenue?
Between	What are the movies that earned revenue between {Min Number} to {Max Number} in revenue?

れる．実データに基づいて作成されたコーパスは Movie Revenue（映画興行収入），Job Post（求人），Company Employees（企業従業員数）の三つである．

Movie Revenue は Movie dataset から作成された映画の興行収入に関するデータセットになっており，映画のタイトルとアクション，サスペンス，SF などの映画のジャンル，その映画の制作会社と興行収入の数値データを含んだものとなっている．また，作成するにあたって対象としたデータは英語の映画タイトルのみとしており，加えて，ジャンルや制作会社，興行収入のデータが存在していないものは除外するようにした．

また，Job Post と Company Employees は LinkedIn dataset から作成した．Job Post は求人情報に関するデータセットになっており，LinkedIn dataset の求人情報と職能情報，業界情報を参照し，会社名，タイトル，説明，収入の数値情報，求人の職能，求人の業界を含んでいる．作成するにあたって，求人情報は年収単位での情報のみに絞り，収入の数値情報はその求人での最高収入の数値を利用した．会社名，タイトル，説明，収入の数値情報を含んでいないものや複数の求人で説明が重複しているものは除外するようにした．Company Employees は会社ごとの従業員数に関するデータセットになっており，LinkedIn dataset の会社情報と業界を参照し，会社名，会社説明，従業員数，会社の業界，本社の場所を含んでいる．作成するにあたって，会社名，会社説明，従業員数，業界，本社の場所を含んでいるものを対象とした．本社の場所については表記揺れ等が存在するため，正規化を行い，その後，場所が同じである会社が 10 件未満のものを除外するようにした．また，会社説明が重複しているものも除外した．

これらの実データセットに基づいて，いくつかの種類のテンプレートを用いてコーパスを作成した．実データの値をテンプレートの変数部分に割り当てることによってコーパスが作成される．また，作成されるコーパスのエンティティと数値の分布は実データと同じである．

加えて，詳細な分析を行うために複数の合成データセットも作成した．これらの合成データセットも同様にテンプレートを用いて作成されている．

Unit データセットは，様々な種類の数値に対する検索モデルの数値処理能力を調査するために作成され，重量や速度，距離などの様々な数値単位を想定したコーパスである．それぞれの数値単位カテゴリは数値と Faker[*1] によって生成したエンティティ名のペアで構成されており，全てのカテゴリでエンティティと数値のペアは同じである．数値の範囲は 1 から 1000 までの 999 通りあり，一つの数値に対して 5 パターンのエンティティ名を作成した．数値単位のカテゴリは 9 つあり，通貨，時間，重さ，長さ，面積，体積，温度，速度，電力である．また，時間というカテゴリであれば分，時，年など複数パターンの単位表現で作成しており，全部で 27 通りの数値単位のデータセットを作成した．

Corporate Value データセットは値のスケールや値の表現方法の違いが検索の有効性に与える影響を調査するために構築した．企業名は Faker によって生成され，業界は LinkedIn データセットを利用した．企業が属する国は 13 の国の中からランダムに選択し，データのペアを作成した．企業価値の数値部分を表す Corporate value は 1-1,000，1-1,000k，1-1,000M の計 2997 通りの数値が存在しており，これらの数値は 3 通りの数値スケールの表現，3 通りの数値の単位表現，2 通りの数値の桁表現の組み合わせの 18 通りでコーパスを作成した．数値スケールの表現はそのままの場合（×1），千の単位で省略して表記する場合（×1k），百万の単位で省略して表記する場合（×1M）が存在する．例えば，数値が$1,000,000$ であるとき，“$1,000,000$” と “$1,000$k”, “1M” となる．数値の単位表現はドル，USD，$ で表現する．例えば，数値が$1,000,000$ としたとき，“1,000,000 dollars”, “1,000,000 USD”, “$ 1,000,000” として表現する．数値の桁表現はカンマを含める，含めないかで表現する．例えば，数値が$1,000,000$ であるとき，カンマを含める場合は “$1,000,000$”，含めない場合は “1000000” として表現する．最終的に企業価値の数値部分は，“$1,000$k dollars”, “1000000 USD” というように 3 つの表現パターンを組み合わせて作成される．

TOEFL Score データセットと Job Salary データセットは，RQ3 を調査するために作成したデータセットである．TOEFL Score データセットは TOEFL のスコアに関する合成データセットであ

[*1] https://faker.readthedocs.io/

り，Job Salary データセットは年収に関する合成データセットである．TOEFL Score データセットでは，国籍，性別の違いによって TOEFL スコアにバイアスが存在しているのかについて調査するために作成し，英語が母国語でない国は母国語である国よりも実際の TOEFL スコアよりも低く見積もられる可能性があるのかなどを調査した．Job Salary データセットでは，国籍，性別の違いによって収入にバイアスが存在しているのかについて調査するために作成し，女性が男性よりも実際の年収よりも低く見積もられる可能性があるのかなどを調査した．どちらのコーパスにも人名，性別，業界，スキルを含んでいる．人名は names-dataset[*2]から国別の一般的に利用される名前を取得して利用し，スキル（Skill），業界（Industry）は LinkedIn データセットを利用した．国は 11 通り用意し，名前は男女それぞれで 500 件ずつ取得して利用した．スキルと業界，数値情報のペアをランダムに 5000 件作成し，それを性別，国別それぞれのパターンで作成した．TOEFL Score データセットの数値部分である TOEFL スコアは TOEFL で取りうるスコアである 0 から 120 までの 121 通りがある．一方，Job Salary データセットの数値部分である Salary は 10,000 から 150,000 の範囲を対象で数値の間隔を 500 に設定し 100 通りの数値がある．

実験で使用したすべてのクエリには数値条件が含まれている．文書中の数値を v_d，クエリ中の数値を v_q とすると，使用した 5 つの数値条件は以下のように定義される：等価（Equal）（$v_d = v_q$），以下（Less）（$v_d \leq v_q$），以上（More）（$v_d \geq v_q$），近似（Around）（$0.85 \times v_q \leq v_d \leq 1.15 \times v_q$），範囲（Between）（$v_q \leq v_d \leq v'_q$）．これらの条件は，クエリとして使用される際に自然言語で表現される．クエリもコーパスの作成方法と同様に数値条件ごとのテンプレートをもとに作成した．例えば Movie Revenue では表 2に示すテンプレートを各数値条件クエリの作成で利用した．

テンプレートの使用例として，例えばv_q = 7630000 の等価（Equal）クエリの場合は，“What are the movies that earned equal to 7630000 dollars in revenue?” といったクエリになる．特に指定がない限り，クエリ中の数値はコーパス内の数値からランダムに選択した．各文書の真の数値情報にアクセスできるため，各クエリに対して関連文書の集合を正確に定義することができる．また，範囲（Between）クエリの場合は 2 つの数値をランダムに選択し，その数値を利用した．

加えて，TOEFL Score や Job Salary データセットでのクエリについては数値条件だけでなく，エンティティも含めた検索を行う場合がある．エンティティ検索も含めた数値条件検索クエリは，“Who has a TOEFL score of 100 points among those who work in Product Management jobs in Entertainment companies?” というようなクエリとなる．TOEFL Score, Job Salary でエンティティ検索として対象となるのはスキル（Skill）と業界（Industry）であり，両方のエンティティの検索を行う場合と片方のエンティティ検索のみを行う場合がある．

本研究で使用した検索モデルは，BM25，DPR（bert-base-uncased を基盤モデルとし，Natural Questions [12] で微調整），E5（E5-base[*3]および E5-large[*4]），ColBERTv2[*5]，および OpenAI の埋め込み [14]（text-embedding-3-small[*6]）である．

数値条件クエリにおいて関連文書が多く存在する可能性を考慮し，主要な評価指標として nDCG@100 を選択した．本データセットでは各クエリに複数の正例が存在し，とりわけ Less/More では正例が多数となり得るため，検索上位から一定深さまでの結果を対象に全体的な傾向を把握できる指標として nDCG@100 が適切と判断した．

この実験設計により，実データと合成データを用いて，数値条件を含むクエリに対する様々な検索モデルの性能を包括的に評価することができる．実データを使用することで実践的なシナリオにおける検索性能への洞察が得られ，合成データにより多様なクエリパターンにわたる徹底的な検証が可能となる．複数の検索モデルを比較することで，数値クエリの処理における各アプローチの特性，長所，短所を明らかにする．

3.2 RQ1: 密検索モデルは NumQ に対して有効に働くか？

図 1は，3 つのクエリ条件における各検索モデルの検索性能を示している．（1）数値条件のない **NumQ**，（2）**NumQ**，（3）数値条件による事後フィルタリングを適用した **NumQ**．条件（1）は **NumQ** と同じような文章で数値条件部分の表現をなくしたクエリを作成し，それを利用した．条件（3）では，各検索結果の真の数値情報を参照して，各検索モデルの検索結果において無関係な結果をフィルタリングした．したがって，（3）は各検索モデルが数値条件を完全に考慮する場合をシミュレーションしている．（1）と（2）の差異は，各密検索モデルが **NumQ** の数値条件をどの程度適切に処理できるかを示し，（2）と（3）の差異は改善の余地を示している．

Tukey’s HSD 検定の結果，Movie Revenue データセットの DPR を除き，（1）-（2）と（2）-（3）の間に有意な差が認められた．ほとんどの密検索モデルにおいて，（2）の検索性能は（1）よりも高く，これは密検索モデルが **NumQ** の数値条件を認識していることを意味する．しかし，（2）と（3）の間には大きな差があり，**NumQ** に対する密検索モデルの能力が低いことを示している．

また，比較した検索モデルの中で，ColBERT は比較的高い検索性能を示した．これは，ColBERT がトークンレベルでの類似度検索を行っており，より念密な検索を可能にしているためと考えられる．この傾向は，[1] の知見と一致している．加えて，単一ベクトルでの検索を行う DPR，E5，OpenAI では，各データセットで性能の高いモデルが異なり，データセットで扱う内容や数値表現方法の違いによって性能に変化があること推察される．

3.3 RQ2: どのような種類の NumQ に対して有効に働くか？

本節では，密検索モデルがどのような種類の **NumQ** に対して有効に働くのかを多角的に検証した．具体的には，数値条件の種類（等価，以下，以上，近似，範囲），数値の単位（時間，重量，長さなど），数値の表現方法（”5000 dollars”，”$5000”，”5k dollars”

[*2] https://pypi.org/project/names-dataset/

[*3] https://huggingface.co/intfloat/e5-base-v2

[*4] https://huggingface.co/intfloat/e5-large-v2

[*5] https://huggingface.co/colbert-ir/colbertv2.0

[*6] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

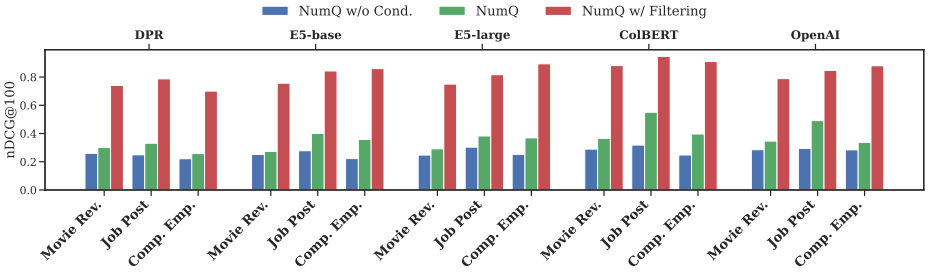

図1　3 つのクエリ条件における各密検索モデルの検索性能

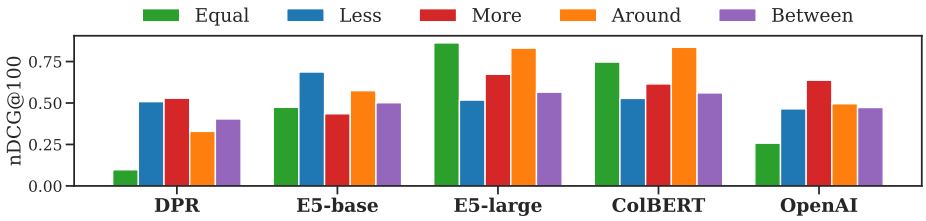

図2　各数値条件における密検索モデルの検索性能

など）の 3 つの側面から，各密検索モデルの検索性能を分析した．これらの分析を通じて，密検索モデルが **NumQ** を処理する際の強みと弱みを明らかにし，今後の改善に向けた知見を得ることを目的としている．

3.3.1　数値条件

図 2は，Unit データセットにおける異なる数値条件に対する各密検索モデルの検索性能を示している．Tukey's HSD 検定（$\alpha = 0.05$）により各密検索モデルにおける数値条件の違いによる検索性能の差異は統計的に有意であることが確認された．

以下（Less）・以上（More）の数値条件クエリでの適合文書数の分布はどちらも一様であるのにも関わらず，どの密検索モデルにおいても検索性能に差が生じている．E5-large, ColBERT, OpenAI では More クエリが Less クエリよりも性能が良く，E5-base は Less クエリの方が性能が良い．この結果は，数値条件の違いによって検索の有効性が変化することを強く示唆している．

また，比較的効果的な検索モデルである ColBERT と E5-large は，等価（Equal）および近似（Around）クエリに対して良好な性能を示したが，以下（Less），以上（More），範囲（Between）クエリに対する性能は限定的であった．この結果は，これらのモデルが数値の類似性を測定できるものの，数値の大小関係を完全に認識する能力は不十分であることを示唆している．一方，ColBERT と E5-large と比較して，DPR と E5-base，OpenAI は等価（Equal）クエリに対して有意に低い検索性能を示した．この結果から，これらのモデルは数値を正確に理解しておらず，抽象的に数値の値を考慮して検索している可能性を示唆している．

3.3.2　数値表現

図 3は，企業価値データセットにおける異なる数値表現に対する各モデルの検索性能を示している．一元配置分散分析（one-way ANOVA）検定を各モデルに対して実施した結果，E5-large，および ColBERT において，数値のスケール表現，数値の単位表現，数値の桁表現の全ての数値表現間に統計的に有意な差が認められた（$\alpha = 0.05$）．DPR，E5-base，OpenAI は一部の数値表現間に統計的に優位な差が認められなかった．

この結果から，本研究で検証を行ったモデルの中でも **NumQ** に対して高い性能を示した E5-large や ColBERT については，数値に対して高い理解力があるために返って優位差が認められたと考えられる．そして DPR や E5-base については，数値に対しての理解力が高くないため，多少の表現の違いがあっても区別されず，結果として表現間で優位な差が見られなかったと考えられる．

特に，ColBERT においては，M（百万を意味する略語）は他の表現よりも低い検索性能で，USD も dollar などの他の表現よりも低い検索性能を示した．これは一般的ではない数値の表現では検索性能が低下する懸念があることを示唆している．

3.3.3　数値単位

図 4は，Unit データセットにおける異なる数値単位カテゴリに対する各モデルの検索性能を示している．二元配置分散分析（two-way ANOVA）検定を各モデルに対して実施した結果，全てのモデルにおいて，異なる数値単位間に統計的に優位な差が認められた（$\alpha = 0.05$）．この結果は，密検索モデルは数値単位の違いに応じて，検索性能が変化することを示している．この実験では 9 つの数値単位カテゴリで検証を行ったが，一般的でない数値単位の表現を行った場合に検索性能が変動する可能性を示唆している．

また，今回検証を行った各モデルの中で ColBERT は F 値が小さく，ColBERT が他のモデルよりも頑健性が高いことが示唆され，DPR や E5，OpenAI などの単一のベクトルでの検索ではなく，トークンレベルのマルチベクトル検索などのより表現能力の高い検索モデルが頑健性が高いことが推察される．加えて，単一のベクトルでの検索では，訓練データセットのドメインの偏りによって **NumQ** に対する検索性能が変動する可能性を示している．

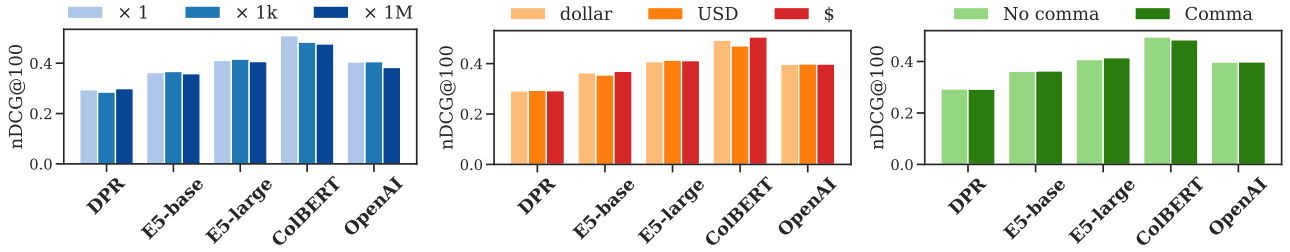

図3　各数値表現における検索性能

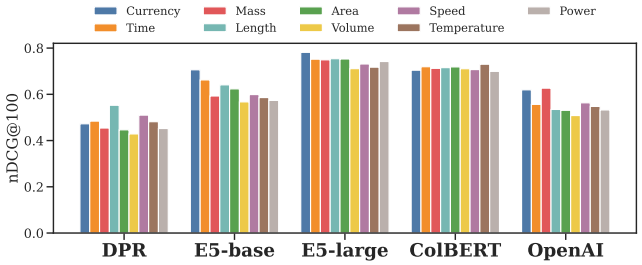

図4　各数値単位における検索性能

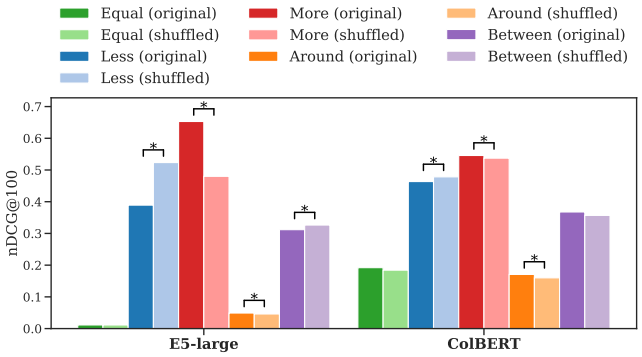

図5　数値がシャッフルされた Movie Revenue データセットとオリジナルのデータセットでの検索性能の比較

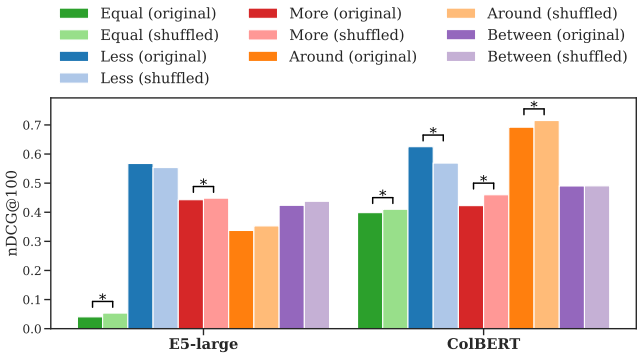

図6　数値がシャッフルされた Job Post データセットとオリジナルデータセットでの検索性能の比較

3.4　RQ3: 言語モデルの内部知識は NumQ の検索に影響を及ぼすのか？

本節では，密検索モデルの学習に用いられる言語モデルが持つ内部知識が，**NumQ** の検索にどのような影響を与えるかを調査した．具体的には，映画タイトルなどの固有名詞が関連する **NumQ** に与える影響，一般常識が **NumQ** に与える影響，センシティブな属性（国籍や性別）が資格スコアや年収に関する **NumQ** の検索に与える影響を分析した．また，密検索モデルの基盤モデルを変更した場合に，**NumQ** の検索性能がどのように変化するかについても検証を行った．これらの分析を通して，言語モデルが持つ知識が **NumQ** の検索に及ぼす影響を明らかにし，より公平で正確な検索を実現するための課題を考察する．

3.4.1　固有名詞

“アバターは世界でもっとも興行収入の多い映画である” というような固有名詞と数値の間にある知識が検索に対して影響を及ぼすのかを調査した．密検索モデルの固有名詞に対する知識の影響を調査するため，映画の興行収入データセットの映画のタイトルなどのエンティティと興行収入の数値情報のペアをシャッフルし，元のデータセットとの性能差を調査し，その結果を図 5に示す．シャッフルされたデータセットでは数値の分布に変更はないため，同一の **NumQ** に対して性能差は生じないことが期待され，有意な性能差が認められた場合には，密検索モデルが与えられた **NumQ** 内の数値条件に関連する文章を特定する際に，映画のタイトルなどの非数値情報に大きく依存していることを示している．

各モデルの元のデータセットとシャッフルされたデータセットに対して対応のあるt 検定を実施し，p 値に Holm 補正を適用した．統計的に有意な差（$\alpha = 0.05$）は，図に$*$ で示されている．

シャッフル後，特に E5-large は以下（Less）クエリに対してより高い性能を示す一方，以上（More）クエリに対してはより低い性能を示した．これは，**NumQ** が興行収入に関する数値条件のみを含んでいたにもかかわらず，E5-large が映画タイトルに基づいて関連性を推定していたことを示している．以下（Less）クエリの性能がシャッフルされたデータセットで向上したことから，E5-large は高い興行収入を得ている主要な映画に対して高いスコアを与える傾向があることが推測される．一方，ColBERT は E5-large よりも小さな性能差を示しており，E5-large よりも固有名詞に対する影響が小さいことが推測される．

3.4.2　一般常識

“医者は比較的収入が高い”，“経営層のポジションは比較的収入が高い” という一般常識が **NumQ** の検索に対して影響を及ぼすのかを調査した．密検索モデルの一般常識に対する知識の影響を調査するため，求人データセットのエンティティ情報とその求人の収入の数値情報のペアをシャッフルし，元のデータセットとの性能差を調査し，その結果を図 6に示す．前節の固有名詞の節

と同様に，シャッフルされたデータセットでは数値の分布に変更はないため，同一の **NumQ** に対して性能差は生じないことが期待される．優位な性能差が認められた場合には，密検索モデルが与えられた **NumQ** 内の数値条件に関連する文章を特定する際に，一般常識が影響を及ぼしていることを示している．

前節（3.4.1）と同様の検定手順に従って統計解析を行った．統計的に有意な差（$\alpha = 0.05$）は，図に$*$で示されている．

実験結果から，E5-large，ColBERT の両者で一部の数値条件においてオリジナルとシャッフルされたデータセットの間で優位な差があることが示された．E5-large では等価（Equal），以上（More）において優位な差が示され, ColBERT では範囲（Between）を除く 4 つの数値条件において優位な差が見られた．ColBERT は E5-large に比べ，多くの数値条件で性能に優位な差が見られ，ColBERT は特に一般常識が **NumQ** の検索において影響を及ぼしていることが示唆している．

3.4.3 センシティブ属性

国籍や性別などのセンシティブ属性を含むコーパスを用いて，数値条件を含む検索における公平性を検証した. 例えば, “TOEFL のスコアが 100 点以上の人” というクエリに対して，同じ点数であるにも関わらず，国籍等の違いによって検索の順位に差が生じるといった問題の有無を調査した．TOEFL と収入のデータセットそれぞれにおいて，特定の数値条件と数値の範囲のクエリで検索して取得されたコーパスの数値の分布を性別と国別で分析し，その結果の一部を図 7に示す．横軸は取得されたコーパスの値を示している．検索結果の対象は上位 500 件を対象とした．全体を通して，性別や国別のみで検索されやすさに違いがあり，本研究では，数値条件とセンシティブ属性との関係による検索されやすさの違いについての調査に焦点をあてるため，性別，各国での検索結果集計を正規化した分布をプロットしている．

実験の結果，両データセットで性別・国別に分布の傾向差が確認された．例えば，E5-large で 60k-70k の範囲を指定したクエリでは，ドイツは 85k ドル付近，アメリカは 70k ドル付近，ブラジルは 55k ドル付近の文書が特に取得されやすかった．クエリに含まれるのは 60k-70k の範囲の数値と職能，業種の情報のみで，国名を含まないにもかかわらずこの差が生じたのは，言語モデルの内部知識や通貨の違いによって数値理解に齟齬が生じている可能性が考えられる．

加えて，全体的に E5-large は性別，国別での傾向の違いが大きく，ColBERT は傾向の違いは小さかった．これは，ColBERT が各トークン別を対象にして類似度を計算しているため，より数値部分に焦点を当てた類似度の推論が可能であるためと考えられる．一方，E5-large は全てのトークンを加味して単一のベクトルに変換するため，数値以外の情報が影響を及ぼしやすいと考えられる．

これらの結果は，密検索モデルにおいて，国籍や性別などのセンシティブ属性の情報が **NumQ** の検索に影響を及ぼし，公平性が必要とされる **NumQ** の検索ニーズに対して十分な性能を示すことができないことが推測される．

3.4.4 基盤モデル

BERT に加えて，より高い数値処理能力を獲得するために事前学習された複数の言語モデルで密検索モデルを作成し，その検索性能を分析した．FinBERT [2]，SEC-BERT（SEC-BERT および SEC-BERT-NUM）[13]，および GenBERT [6]．FinBERT は金融文書を用いて追加の事前学習が行われている．SEC-BERT は財務文書を用いて追加の事前学習が行われており，SEC-BERT-NUM は，数値部分のトークンの処理方法に改良を加えて学習されたものである．GenBERT は, 数値推論に特化した大量の合成データで追加の事前学習が行われている．これらの言語モデルを基盤モデルとして, DPR モデルを作成した．DPR は Natural Questions [12] データセットを用いて学習された．

図 8は，異なる言語モデルを用いた DPR の性能を示している．二元配置分散分析の結果，すべてのデータセットにおいて有意な差が認められ, Tukey’s HSD 検定では（BERT, SEC-BERT-NUM）と（SEC-BERT, GenBERT）のペアを除くすべてのモデルペア間で有意な差が示された（$\alpha = 0.05$）．

特筆すべきは，SEC-BERT と GenBERT が通常の BERT に対して安定した改善を達成したことである．しかし，これらの性能は図 1に示された理想的な性能にはまだ遠く及ばない．したがって，密検索モデルが NumQ を正確に処理するためには，抜本的な解決策が必要だと考えられる．

4 まとめ

本研究では，密検索モデルが数値条件を含むクエリ（**NumQ**）に対してどの程度有効であるかを DPR, E5, ColBERT などの主要な密検索モデルを用い，実データと合成データを組み合わせた複数のジャンルのデータセットで詳細に調査した．まず，RQ1 では，密検索モデルが **NumQ** に対してどの程度有効であるかを検証した．実験の結果, 密検索モデルはある程度 **NumQ** の数値条件を考慮した検索を行えることが示されたものの，理想的な性能には大きな乖離があり，改善の余地が大きいことが明らかになった．具体的には，数値条件を含まないクエリよりも数値条件を含んだクエリに対する検索性能は高いものの，数値条件を完全に満たす検索結果と比較すると大きな乖離があった．この結果は，密検索モデルが数値条件を認識はできるものの，その能力はまだ十分ではないことを示唆している．次に，RQ2 では，どのような種類の **NumQ** に対して密検索モデルが有効に働くのかを詳細に分析した．数値条件の種類（等価，以下，以上，近似，範囲）の違いや，時間や重量，長さなどの数値単位の違い，数値表現の違い（例：「5000 dollars」,「$5000」,「5k dollars」）が検索性能に与える影響を調査した．実験の結果，数値条件の種類や数値表現，数値単位の違いによって，密検索モデルの検索性能が異なることが明らかになった．特に，等価（Equal）や近似（Around）クエリでは良好な性能を示した一方, 以下（Less), 以上（More), 範囲（Between）クエリでは性能が限定的であることが分かった．この結果は，密検索モデルが数値の類似性を測定できるものの，数値の大小関係を完全に認識する能力は不十分であることを示唆している．また，数値の単位や数値表現が異なると性能が変動することも明らかになった．RQ3 では，言語モデルの内部知識が **NumQ** の検索に与

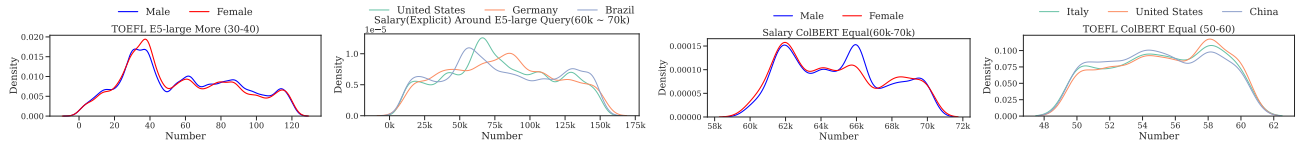

図7　TOEFL，収入データセットにおける検索結果上位 500 件のコーパスの数値を性別および国別で正規化された頻度分布

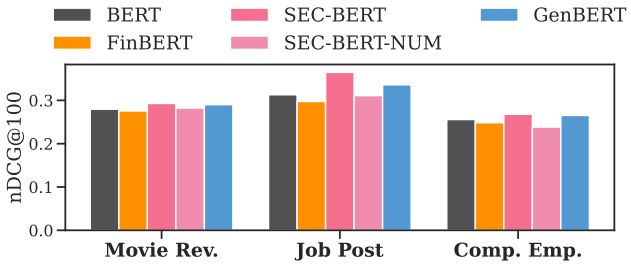

図8　各基盤モデルから作成した DPR の各データセットにおける検索性能

える影響を調査した．具体的には，映画タイトルなどの固有名詞や一般常識が **NumQ** の検索に与える影響，国籍や性別などのセンシティブな属性が資格スコアや年収に関する **NumQ** に与える影響を分析した．実験の結果，言語モデルの事前学習で獲得された固有名詞や一般常識が **NumQ** の検索結果に影響を及ぼし，偏った検索結果につながる可能性があることが分かった．また，密検索モデルの基盤モデルを数値に関連したデータセットや数値に特化した学習手法によって学習された言語モデルに変更することで，検索性能に変化があることも明らかになった．しかし，これらの改善策によっても，理想的な性能とは未だに大きな隔たりがあることが示された．これらの知見は，数値条件を含むクエリに対する検索システムの設計・開発において重要な示唆を与える．今後の研究では，数値条件をより正確に理解し，バイアスの影響を軽減する手法の開発が課題となる．さらに，本研究で用いたデータセットや評価指標の限界を踏まえ，より多様なデータセットや評価指標を用いた検証も必要である．

謝辞

本研究は JSPS 科研費 JP24K03048, JP23K28090, JP23K25159 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Satya Almasian, Milena Bruseva, and Michael Gertz. Numbers matter! bringing quantity-awareness to retrieval systems. *arXiv preprint arXiv:2407.10283*, 2024.
[2] D Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
[3] Rounak Banik. The movies dataset. `https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset`, 2017. Accessed: 2024-10-15.
[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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
[5] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*, pages 2368–2378, 2019.
[6] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. In *ACL*, pages 946–958, 2020.
[7] Vinh Thinh Ho, Yusra Ibrahim, Koninika Pal, Klaus Berberich, and Gerhard Weikum. Qsearch: Answering quantity queries from text. In *ISWC*, pages 237–257, 2019.
[8] Vinh Thinh Ho, Koninika Pal, and Gerhard Weikum. Qute: Answering quantity queries from web tables. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2740–2744, New York, NY, USA, 2021. Association for Computing Machinery.
[9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.
[10] 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*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery.
[11] Arsh Kon and Zoey Yu Zou. Linkedin job postings (2023 - 2024). `https://www.kaggle.com/datasets/arshkon/linkedin-job-postings`, 2023. Accessed: 2024-10-15.
[12] 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*, 7:453–466, 08 2019.
[13] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Paliouras George. FiNER: Financial numeric entity recognition for xbrl tagging. In *ACL*, 2022.
[14] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training, 2022.
[15] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics.
[16] 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'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
[17] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *EMNLP-IJCNLP*, pages 5307–5315, 2019.
[18] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024.