DEIM Forum 2023 3a-2-3

単一密検索モデルによる複数言語横断情報検索

阿部　健也† 　新田　洸平†† 　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
†† 筑波大学大学院 人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1–2
††† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{s1911448,s2221648}@s.tsukuba.ac.jp, ††mpkato@acm.org

あらまし　本研究では，単一の密検索モデルによって，複数の言語間の言語横断情報検索 (CLIR) を実現する方法を提案する．従来の密検索を用いた CLIR では，各言語毎に検索モデルを用意するアプローチが用いられることが多く，精度が高く有効ではあるものの，言語ごとにモデルを管理する必要があること，また，データを用意して学習を行う必要があることが課題であった．我々はこの問題を解決するために，1 つの事前学習多言語モデルを採用し，すべての言語のデータを用いた学習を行う方法を試行した．実験では，CLIR のためのテストコレクションを用いて提案手法の単一モデルによる検索の精度を検証し，言語ごとに本提案手法の有効性が大きく変わることを示した．

キーワード　検索モデル、言語モデル、アドホック検索、言語横断情報検索

1 はじめに

近年，言語横断情報検索が注目を集めている．入力としてクエリを与え，出力をクエリに対する適合度が高い文書のランキングとするものを「アドホック検索」と呼び，その中でもクエリの言語と文書の言語が異なっている場合を「言語横断情報検索 (Cross Language Information Retrieval，以下 CLIR) 」と呼ぶ．この技術は母国語以外の言語で書かれた特許文献を検索したいと考えた時や，海外のローカルなニュース記事を検索したい時に必要になる．CLIR は古くから取り組まれてきた問題で近年注目が集まっている問題でもある．情報検索のワークショップである TREC [17] や NTCIR [14] においては 1990 年代から開催されていて，2022 年にも TREC において NeuCLIR という CLIR のタスクが開催された．

近年，事前学習多言語モデルを利用した密検索モデルの CLIR への適用が進んでいる．CLIR への適用に注目が集まる以前から，単言語のアドホック検索においては密検索モデルが有効に作用することがわかっていた．密検索とは BERT [5] などの事前学習言語モデルを利用して単語をベクトルに変換し，類似度を計算する手法である．事前学習言語モデルとは，モデルを事前に大量の文章データで学習させることで単語の文脈的意味を表現することができるモデルである．このモデルを利用することにより密検索は単語の意味まで考慮できるため，BM25 などの単語一致の手法よりも効果的である．この密検索モデルの特徴は，クエリと文書の言語が異なっている CLIR においても有効に作用すると考えられ，近年研究が行われている [12] [16] [15] [19]．CLIR では，クエリと文書の言語が異なっているため BM25 などの手法を直接適用することができない．そのため，単語一致の手法で CLIR に取り組む際は，クエリか文書のどちらかを翻訳して言語を一致させて検索を行う．この手法の課題として，検索の精度が翻訳に依存してしまうため，翻訳した結果が間違っていると目的とする文書を検索することができなくなってしまうことが挙げられる．この課題に対して，単語の意味を考慮できるため検索する際に翻訳を必要としない事前学習多言語モデルを利用した密検索が有効であると考えられる．事前学習多言語モデルとは事前学習の際に様々な言語の文章データを学習したモデルのことである．このモデルは同じ意味を持っている異なる言語の単語を似た意味の単語として表現することができるため翻訳を必要としない．具体的には，「cat」という単語で日本語の文書検索をすると，単語一致の手法では「cat」を「猫」に翻訳するという過程が必要になる．しかし，事前学習多言語モデルは「cat」と「猫」が同じ意味であることを表現できるため翻訳をする必要がない．よって，事前学習多言語モデルを利用した手法である密検索は CLIR に効果的に作用すると考えられる．実際に ColBERT-X [12] では，事前学習多言語モデルを利用して CLIR に取り組む手法を提案し，クエリと文書の言語を一致させるために翻訳を必要とする BM25 よりも高い精度が得られた．ColBERT-X では，英語の学習データで 1 つのモデルを学習させてそのモデルで各言語の検索を行う手法 (Zero-Shot) と，英語の学習データを翻訳することで言語毎の学習データを生成し，各モデルでその言語に特化するような学習をして検索を行う手法 (Translate-Train) の 2 つを提案し，2 つの提案手法の比較では Translate-Train がより有効であることを示した．

言語毎に複数のモデルを用意する手法である Translate-Train は精度が高く有効ではあるものの，複数のモデルを管理する必要があること，モデル毎に学習を行う必要があることが課題である．我々は，事前学習多言語モデルが目的の言語以外の学習データを利用しても目的の言語でのタスクの精度が向上するという特徴に注目した．さらに Asai ら [1] は文章生成を行う際の事前学習多言語モデルのファインチューニングに複数の言語が混ざった学習データを利用しても文章生成の精度が向上することを示した．これらのことから，文書検索においても複数の言

語の学習データをまとめて 1 つの学習データとして利用することで複数モデルを用意しなくても検索の精度が向上することを期待する．よって，本研究では単一のモデルで目的とする言語毎により精度の高い検索が行えるような手法として，1 つの事前学習多言語モデルを用意し，各言語毎の学習データを 1 つの学習データとしてまとめて学習させる手法を提案する．

実験では，CLIR のためのテストコレクションである HC4 [11] というデータセットで，提案手法やベースライン手法による検索を行った．ベースライン手法として BM25 や事前学習多言語モデルを英語のデータのみで学習させたモデル (Zero-Shot) による検索，目的の言語の学習データで学習させたモデル (Translate-Train) による検索と提案手法による検索とリランキングを行い，これらの結果を比較した．リランキングとは他の手法によって生成された検索結果 (実験では BM25 による検索結果) を並び替える手法を呼ぶ．また，我々は TREC 2022 で開催された NeuCLIR というトラックに参加し，提供されたデータセット neuclir1 に対して提案手法を適用した結果を提出した．NeuCLIR ではベースラインとして文書を翻訳することでクエリと文書の言語を一致させ，BM25 による検索を行った結果が与えられているため，その結果と提案手法の結果を比較した．

HC4 に対する実験で，中国語では BM25 の結果と比較して提案手法が統計的に有意に上回ったが，Translate-Train の結果を上回ることはできなかった．ペルシャ語では BM25 と比較して提案手法が統計的に有意に上回り，他の手法の結果も上回ったが有意差は認められなかった．ロシア語では提案手法が BM25 の結果を下回り，全ての手法の間で統計的に有意な差は認められなかった．neuclir1 に対する実験結果から，提案手法による全文書検索では，全ての言語の Recall@1000 において提案手法がベースラインの BM25 を統計的に有意に下回った．Recall@1000 の値が下回ったにも関わらず，ロシア語では他の指標が BM25 を統計的に有意に上回った．また，提案手法によってベースラインをリランキングした結果が全ての言語の Recall@1000 以外の指標においてベースラインを上回り，いくつかの指標では統計的な有意差も認められている．以上の結果から，neuclir1 において提案手法はリランキングに用いると特に有効に作用すること，提案手法の有効性は言語のみに限らずデータセットによっても異なることがわかった．本論文の貢献は以下の通りである:

（1） 事前学習多言語モデルに基づく CLIR において，複数の言語の CLIR に取り組む際，複数の言語の学習データをまとめて 1 つのモデルで学習する手法を提案した．提案手法では，1 つのモデルで複数の言語の CLIR に対応することができる．

（2） 2 つのデータセットに対する実験から，提案手法が言語によって有効に作用することを示した．また，提案手法の有効性が言語のみではなく，データセットによっても異なることを示した．

本論文の構成は以下の通りである．2 節では事前学習言語多言語モデル，事前学習多言語モデルによる CLIR に関する既存研究について述べる．3 節では提案手法の概要を説明し，ColBERT，ColBERT-X，および提案手法の詳細について述べる．4 節ではデータセットと実験の設定，実験結果を示し，5 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，事前学習言語モデル，および，事前学習多言語モデルについて述べたのち，言語横断情報検索への密検索モデルの適用に関する既存研究について述べる．

2.1 事前学習多言語モデル

本節では，事前学習言語モデルと事前学習多言語モデルについて述べる．自然言語処理のタスクを解くために事前に大量の文章データで学習させたものを事前学習言語モデル [5] と呼ぶ．特に様々な言語に対応するために，事前学習の際に複数の言語のデータで学習を行っている場合のモデルを事前学習多言語モデルと呼ぶ [3]．事前学習では大量の文章データから，単語の文脈的意味や単語同士の関係を表すベクトル空間を学習する．その結果，事前学習言語モデルでは様々なタスクを解く際に，単語をベクトルに変換することで単語の意味を考慮できるようになる．事前学習言語モデルでは追加の学習を行うことで文書分類やアドホック検索など，目的に応じた自然言語処理のタスクを解くことができるようになる．このように，既に学習済みのモデルを利用して，目的のタスクのためにそのモデルを追加で学習させることをファインチューニングという．また，事前学習多言語モデルでは事前学習時に含まれている言語であれば，目的の言語ではない言語の学習データをファインチューニングに使用してもタスクの結果が向上するという特徴がある．例えば，文書分類を行うために事前学習多言語モデルを英語の学習データでファインチューニングしても，事前学習時に日本語のデータが含まれていれば，日本語の文書分類の結果が向上する．この特徴は目的の言語においてタスクに対応するための学習データが十分でない場合において非常に有効である．

2.2 言語横断情報検索への密検索モデルの適用

本節では，言語横断情報検索への密検索モデルの適用に関する関連研究について述べる．これまでの単言語のアドホック検索では，BM25 やクエリ尤度モデルなどの単語一致を利用した手法が長い間主流となっていた．単語一致を利用する手法では，同義語や言い換え表現などを関係のない単語として見逃してしまうことが課題とされる．BERT をはじめとした事前学習言語モデルが登場し，単語をベクトルとして数値化することで単語の意味まで考慮して検索を行う密検索手法 [8, 9, 18, 20] が提案されたことで同義語や言い換え表現の意味をある程度考慮することができるようになった．密検索は単語一致の手法よりも有効であることが分かっているため，近年密検索手法が主流となってきている．密検索手法は計算の量が非常に多く全てのクエリと文書のペアに対して類似度を計算しようとすると膨大な時間がかかってしまう．よって，計算時間削減のためにリランキングという手法がとられることがある．リランキングとは，早く動作する BM25 などの手法によってクエリに適合する文書

の候補を全体の文書から決められた数 (多くは上位 1000 件) 検索する．その後，密検索手法でクエリと検索した文書それぞれの類似度を計算し並び替えを行うことで最終的なランキングを決定する．例えば，Yilmaz らが提案した，Birch [20] では，クエリと文書のペアを BERT に入力し，そのペアの適合度を出力する．この手法では，ペアの数だけ BERT による計算が必要になり，大きな計算コストがかかるため，BM25 によって出力されたランキングをリランキングする．一方で，Karpukhin らが提案した Dense Passage Retrieval (DPR) [8] は，クエリと文書を別々に BERT に入力してベクトルに変換し，その 2 つのベクトルの内積により類似度を計算する手法である．この手法は，クエリと文書を別々に変換すること，ベクトルを高速で検索できる近似最近傍探索 (Approximate Nearest Neighbours) [7] を利用していることから，計算が高速で，全文書の検索が可能である．Khattab らが提案した ColBERT [9] では，BERT によってクエリと文書をトークン毎にベクトルに変換してから類似度を計算する．また，リランキングを行うために全文書から検索する手法 (1 段階目) と 1 段階目で抽出した結果をリランキングするための手法 (2 段階目) を提案している．この時，1 段階目と 2 段階目を続けて行う場合を全文書検索，他の手法によって候補の文書を選び，2 段階目のみを行う場合をリランキングと呼ぶ．

CLIR においても同様に，単語一致を利用した手法が主流であったが，近年では密検索手法が注目されている．CLIR では，単言語のアドホック検索において有効とされる手法を CLIR に適用するために改良されることが多い．Shi らは単言語のアドホック検索において有効である DPR を CLIR に適用する手法 (mDPR) を提案した [16]．mDPR では，DPR で利用している事前学習言語モデル (BERT) の部分を事前学習多言語モデルである multilingual BERT (mBERT) [10] に置き換えてファインチューニングすることで CLIR に取り組む．Nair らは，ColBERT を CLIR に適用した手法である ColBERT-X を提案した [12]．ColBERT-X では利用する言語モデルを BERT から事前学習多言語モデルに置き換えるが，mBERT ではなく XLM-RoBERTa [3] に置き換えている．ColBERT-X では，CLIR に適用するための工夫として言語毎にモデルを用意して，モデル毎に言語に特化する学習を行う手法を提案している．この手法は，HC4 [11] や CLEF [2] などのいくつかのデータセットにおいてこれまで報告されていた中で最も良い結果を残した．

本研究の提案手法は ColBERT-X を元にしているため，ColBERT-X との我々の研究の違いについて述べる．ColBERT-X では，言語毎にモデルを用意して，モデル毎に言語に特化する学習を行う手法を提案した．しかし，この手法は有効である一方で，モデルを言語毎に管理する必要があること，学習を言語毎に行う必要があることが課題である．1 つのモデルで複数の言語の CLIR を行える Zero-Shot という手法が存在するが，この手法は言語毎に学習を行う手法と比べると精度が低くなってしまう．本研究の提案手法では，複数の言語の学習データをまとめて 1 つの学習データとして利用することで，1 つのモデルで複数の言語の CLIR に対応する．1 つのモデルで複数の言語の CLIR に対応するだけでなく，目的の言語ではない他の言語の学習データをファインチューニングに使用してもタスクの結果が向上するという事前学習多言語モデルの特徴による効果と Translate-Train のように目的の言語の学習データから得られる効果の 2 つが検索の精度を高めることを期待する．

3 手　　法

本節では，1 つの密検索モデルで複数の言語横断情報検索に取り組むための手法について説明を行う．

3.1 提案手法の概要

本論文では言語の種類の集合を L，文書集合を $\mathcal{D} = \bigcup_{l \in L} D_l$ として，ある言語 $l \in L$ の文書集合 D_l に対して，$l \neq l'$ となるような言語 l' のクエリ $q \in Q_{l'}$ が与えられたとき，q に適合する順に文書をランク付けする問題である．特に，本論文では文書集合 D_l 中の $d \in D_l$ に対して，q との適合度スコア $S_{q,d}$ を推定し，スコアを降順に並べることで文書をランク付けを行う．

CLIR では近年，密検索の適用が進んでいて ColBERT-X ではベースラインとして利用されることが多い BM25 を利用した手法よりも高い精度が得られた．ColBERT-X では単言語の検索において有効である ColBERT という手法を CLIR に適用した．この手法の特徴は，各言語毎に事前学習多言語モデルを用意して，各モデルでその言語に特化させるような学習を行なっているところである．この手法では，言語毎にモデルを用意する必要があること，言語毎に学習を用意する必要があることが課題であるとして，本論文では，言語毎にモデルを用意するのではなく 1 つの事前学習多言語モデルで言語集合 L にある全ての言語の文書の検索に対応するための学習方法について考えた．事前学習多言語モデルでは事前学習のデータに含まれている言語であればファインチューニングの際に目的の言語以外の学習データを利用してもタスクの精度が向上することがわかっている．また，Asai ら [1] は文章生成のために事前学習多言語モデルをファインチューニングする際，複数の言語が混ざった学習データで学習を行う手法を提案し，うまくいくことを示した．このことから，文書検索においても複数の言語が混ざった学習データを活用することで複数モデルを用意しなくても検索の精度が向上すると考えた．我々は，言語の集合 $L = \{l_1, l_2, ...l_m\}$ のそれぞれの言語に対応する，各言語の学習データ $T_{l_1}, T_{l_2}, ...T_{l_m}$ を 1 つの学習データ $\mathcal{T} = \bigcup_{l \in L} T_l$ としてまとめて利用する手法を提案する．

提案手法とモデルを言語毎に用意する手法 (Translate-Train) の比較の図を 1 に示す．以降では ColBERT，ColBERT-X について説明した後で最後に提案手法の詳細を説明する．

3.2 ColBERT

ColBERT について説明する．近年，単言語のアドホック検索では，事前学習言語モデルである BERT によってクエリと文書をベクトルに変換して類似度を計算する手法が有効である．その手法の 1 つである ColBERT は，クエリと文書をそれぞれトークン毎にベクトルに変換してその類似度を計算することで

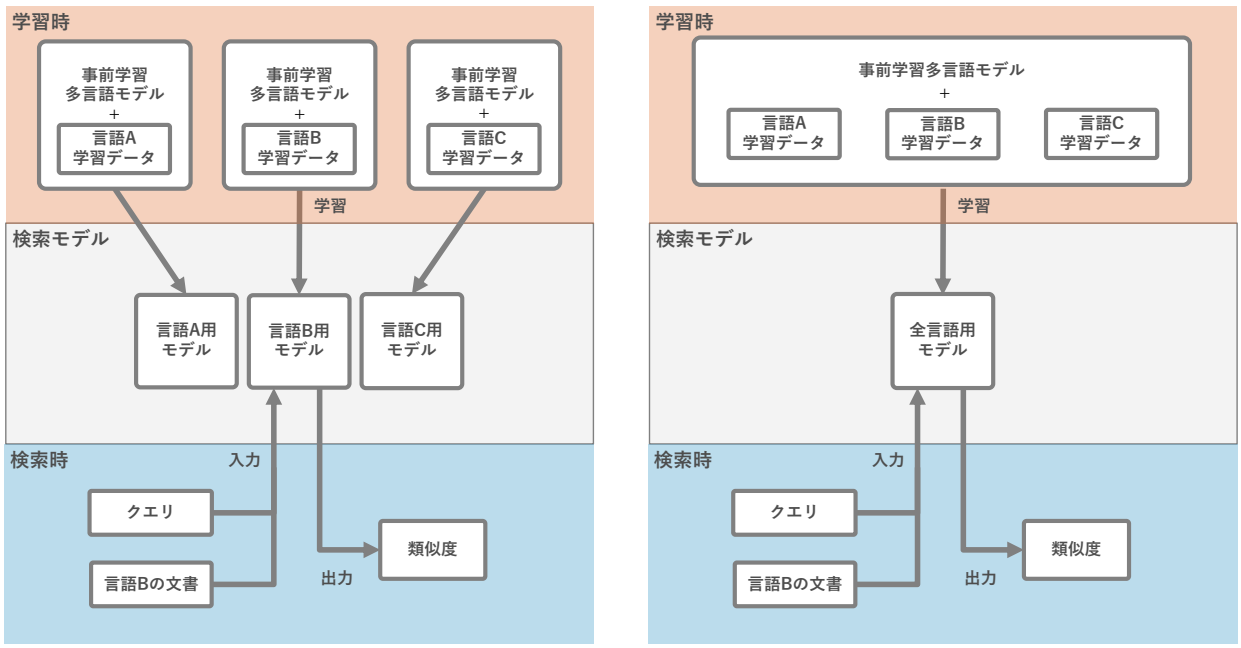

図 1　Translate-Train (左) と提案手法の比較 (右)

検索を行う．ColBERT では計算コストの削減のために，全ての文書の中から並び替えるための候補の文書を選ぶ 1 段階目と候補の文書を上位 1000 件に並び替える 2 段階目の 2 段階に分けた検索を行っている．1 段階目は，近似最近傍探索 [7] を利用して大量の文書から上位の文書になりうる候補の文書を検索する．近似最近傍探索とは，ベクトルの集合に対して，新たにベクトルを与えた時にそのベクトルに近いベクトルを検索するような手法である．この手法は厳密に最も近いベクトルを探すわけではなく，近似的に近いベクトルを探すことで高速化している．近似最近某探索では，文書側のベクトルの集合を用意するために事前に文書の索引付けを行うので，検索の際に必要な計算はクエリのみとなり，検索をより高速で行うことが可能になる．1 段階目を具体的に説明する．文書のトークンのベクトル集合から，クエリのトークンベクトルに類似した文書のトークンベクトルを n 個ずつ近似最近傍探索によって検索する．つまり，クエリのトークンを $|q|$ 個するとクエリトークン 1 つあたり n 個のベクトルを近似最近某探索によって検索するため，$|q| \times n$ 個の文書トークンベクトルを検索することになる．文書のトークンはどの文書に含まれていたものかを対応付けた ID を持っていて，$|q| \times n$ 個選んだ中で，その ID の文書のトークンが多いほどその ID の文書はクエリ q に適合していると考える．$|q| \times n$ 個の中で文書 ID が含まれている数 (重複している数) によって文書を降順に並べ，その上位 k 件を 2 段階目でリランキングする．2 段階目では 1 段階目で抽出したクエリに対する文書のランキングをリランキングする．1 の式はある文書 d のクエリ q に対する類似度を計算する式である．クエリ q，文書 d が与えられた場合，クエリをトークナイザーによってトークンに分割することで $q = (q_1, q_2, \ldots, q_{|q|})$, 文書も同様に分割することで $d = (d_1, d_2, \ldots, d_{|d|})$ とする．これらのトークンは BERT を使用したエンコーダ η によってクエリと文書のベクトル $\eta(q_i)$，$\eta(d_i)$ に変換される．

$$S_{q,d} = \sum_{i=1}^{|q|} \max_{j=1,\ldots,|d|} \eta(q_i)\eta(d_j)^T \tag{1}$$

上の式ではクエリの各トークンのベクトルに対して，文書の各トークンのベクトルとのコサイン類似度を計算し，クエリのトークン毎にそのトークンのベクトルと最も類似度の高い文書のトークンのベクトルの類似度を加算していき，その合計をクエリ q に対する文書 d の適合度 $S_{q,d}$ とする．ColBERT の学習ではあるクエリ q，そのクエリに対する適合文書 d^+，非適合文書 d^- の 3 つで $t_i = (q, d^+, d^-)$ の 1 組 (トリプルと呼ぶ) とした学習データ $T = \{t_1, t_2, \ldots t_{|T|}\}$ を利用する．トリプル t_i をエンコーダに与え，クエリ，適合文書，非適合文書をそれぞれをベクトルに変換した後，クエリ q と適合文書 d^+，クエリ q と適合文書 d^- の 2 つのペアの適合度を計算する．損失関数にクロスエントロピー関数 2 を用いたペアワイズ学習を行う．この手法では事前学習言語モデルのパラメータ集合 $\Theta = \{\theta_1, \theta_2, \ldots \theta_n\}$ を以下の式の 3 を満たすような値に近付けるように学習していくと考えられる．

$$L(q, d^+, d^-) = -\log \frac{e^{S_{q,d^+}}}{e^{S_{q,d^+}} + e^{S_{q,d^-}}} \tag{2}$$

$$\Theta = \arg\min_{\Theta} \sum_{t \in T} L(q, d^+, d^-) \tag{3}$$

ColBERT では検索の処理速度の点から，2 つの手法が用意されている．1 つ目は全文書から候補の文書を選ぶ 1 段階目と候補の文書を並び替える 2 段階目を続けて行う全文書検索という手法である．2 つ目が，別の手法によって得られたランキングを 2 段階目のみを利用して並び替えるという手法でリランキングと呼ばれる．

3.3 ColBERT-X

ColBERT を多言語に拡張した手法である ColBERT-X について説明する．ColBERT-X では，CLIR に取り組むために使用するエンコーダが BERT から，XLM-RoBERTa に変更されている． 事前学習多言語モデルでは BERT を多言語に拡張した mBERT [10] が利用されることが多いが，ColBERT-X ではより良い精度を出すことができることから XLM-RoBERTa を選択している．事前学習多言語モデルはファインチューニングの際に大量の学習データが必要であるため，言語によっては学習データの不足が問題となる．例えば，英語では莫大な量の学習データとして MS MARCO というデータセットが存在するが他の言語において MS MARCO のようなデータセットは存在しない．この問題を解決するために ColBERT-X では Zero-Shot と Translate-Train の 2 つの手法を提案している．1 つ目の手法，Zero-Shot は学習データとして MS MARCO を用いて事前学習多言語モデルをファインチューニングする．この手法では学習データに含まれている言語は英語のみで，他の言語の検索が可能である．これは事前学習多言語モデルの目的の言語とは違う言語を学習データとして利用しても，目的の言語でのタスクの結果が向上するという特徴を利用したものである．また，学習データは英語のクエリと英語の文書で構成されているため，このモデルはクエリと文書の言語が一致している状況での検索を学習する．よって，Zero-Shot では検索時にクエリを文書の言語に機械翻訳して言語を一致させることで検索を行う．Zero-Shot は ColBERT との違いはエンコーダが XLM-RoBERTa になっている点と，検索時にクエリを翻訳するという点である．よって，ある言語 s で表現されたクエリ $q_{(s)}$ とある言語 l で表現された文書 $d_{(l)}$ の類似度 $S_{q,d}$ はクエリを言語 l に翻訳することで $q_{(l)}$ として表すことができ，以下のように表すことができる．

$$S_{q_{(s)},d_{(l)}} = \sum_{i=1}^{|q_{(l)}|} \max_{|d_{(l)}|} \eta(q_{(l)i}) \eta(d_{(l)i})^T \tag{4}$$

2 つ目の手法，Translate-Train は各言語毎に事前学習多言語モデルを用意し，各言語の学習データによってモデル毎の学習を行う．そこで，各言語の学習データは英語のデータである MS MARCO の文書をその言語に機械翻訳することで生成する．MS MARCO は英語のクエリと英語の文書によってデータが構成されている．クエリを別の言語に翻訳することで，新たな学習データは英語のクエリと別の言語に翻訳された文書で構成される．ColBERT から変更されているは学習データの言語が異なっている点であり，類似度の計算や損失関数には ColBERT と同様の式を用いる．よって，クエリの言語を s，検索したい文書の言語を l として，ColBERT でのトリプル t の説明を拡張して考えると，ある言語 l の学習データ T_l を構成するトリプル $t \in T_l$ は $t = (q_{(s)}, d^+_{(l)}, d^-_{(l)})$ で構成される．$q_{(s)}$ はクエリ言語 s で表現されたクエリ，$d^+_{(l)}, d^-_{(l)}$ は検索したい文書の言語 l に翻訳された文書である．これはクエリの言語と文書の言語が異なるという CLIR の問題設定と全く同じである．よって学習後のモデルは，クエリを翻訳することなく CLIR を行うことができる．この時，事前学習多言語モデルのパラメータ集合 $\Theta = \{\theta_1, \theta_2, ...\theta_n\}$ を以下の式 5 を満たすような値に近づけるように学習していくと考えられる．

$$\Theta = \arg\min_{\Theta} \sum_{t \in T_l} L(q_{(s)}, d^+_{(l)}, d^-_{(l)}) \tag{5}$$

3.4 提 案 手 法

提案手法について説明する．本研究の目的は，1 つの事前学習多言語モデルを用意し，ファインチューニングした単一のモデルによって各言語の検索行い，精度を向上することである．我々は，英語の学習データを翻訳して生成した，各言語用の学習データをまとめて 1 つの学習データとして利用する手法を提案する．具体的には，言語の集合 $L = \{L_1, L_2, ...L_m\}$ のそれぞれの言語に対応する，各言語の学習データ $T_{l_1}, T_{l_2}, ...T_{l_m}$ を 1 つの学習データ $\mathcal{T} = \bigcup_{l \in L} T_l$ として利用する．学習の際には，ある言語 s で表現されたクエリ $q_{(s)}$ とある言語 l で表現された $d^+_{(l)}$ と $d^-_{(l)}$ のトリプル $t = (q_{(s)}, d^+_{(l)}, d^-_{(l)})$ $(l \in L)$ が入力される．提案手法は ColBERT に基づいているため，類似度の計算や損失関数は ColBERT と同じ式を用いる．3.3 でのパラメータ集合 Θ の最適化をさらに拡張して考えると，提案手法は事前学習多言語モデルのパラメータ集合 $\Theta = \{\theta_1, \theta_2, ...\theta_n\}$ を下の式 6 を満たすような値に近づけるように学習していくと考えられる．

$$\Theta = \arg\min_{\Theta} \sum_{l \in L} \sum_{t \in T_l} L(q_{(s)}, d^+_{(l)}, d^-_{(l)}) \tag{6}$$

本研究では CLIR のためのテストコレクション HC4，neuclir1 に対して，提案手法を適用しその結果を BM25 や Zero-Shot などのいくつかの手法と比較して検証する．

4 実　　験

本節ではまず使用する学習データと 2 つのテストコレクションとして利用したデータセットについて統計情報を説明する．それからベースライン手法や提案手法の実験設定について述べ，最後に実験結果を示す．

4.1 データセット

表 1 は今回の実験で使用したテストコレクションの詳細で

表 1　テストコレクション HC4 の統計情報

	中国語	ペルシャ語	ロシア語
文書	646K	486K	5.0M
クエリ	60	60	54

表 2　テストコレクション neuclir1 の統計情報

	中国語	ペルシャ語	ロシア語
文書	3.2M	2.2M	4.6M
クエリ	49	46	45

ある．我々は CLIR のためのテストコレクションである HC4 (Common Crawl Collection) というデータセットで実験を行った．HC4 とは Common Crawl の中国語，ロシア語，ペルシャ語のニュース記事から作成されたもので，クエリは全て英語である．文書ではタイトルと本文の他に URL などのいくつかの情報が与えられているが，今回はタイトルと本文のみを利用した．HC4 ではクエリ 1 つあたりに，人手で判定された適合文書がいくつか与えられていてどのくらい適合しているかによって 2 段階に分けられている．1 つ目はクエリによく関連している文書，例えば文書の中心の話題がそのクエリに関するものであるような場合で，2 つ目はクエリにいくらか関連している文書，例えば文書のどこか一部でそのクエリに関する内容が述べられているような場合である．また，適合文書以外は全て非適合文書として扱われる．あるクエリに対しての適合文書は言語によって存在するものもあれば，そうでないものもある．適合文書がない言語ではそのクエリにおいてその言語の文書の検索は行わないためクエリの数は言語によって異なっている．

テストコレクションの 2 つ目は TREC 2022 の NeuCLIR というトラックでテストコレクションとして提供されたデータセットである．クエリと文書の言語や適合度判定は HC4 と同じ構成となっていて，neuclir1 も Common Crawl のニュース記事から作成されたものである．データセットの詳細は表 2 に記載する．

学習データには MS MARCO と neuMARCO を用いた．MS MARCO (a large scale MAchine Reading COmprehension dataset) [13] とはマイクロソフトが公開した英語のデータセットで 50 万のクエリと 880 万の短めの文書が用意されており，クエリ 1 つあたり適合文書が 1 つ以上与えられている．CLIR では，英語以外で大規模な学習データを用意するのが難しいため，各言語の学習データは英語のデータを翻訳して生成される．neuMARCO とは MS MARCO の文書のみを機械翻訳モデル Sockeye [6] で機械翻訳したもので，クエリは MS MARCO と全く同じで，文書は他の言語に機械翻訳されてたものとなる，利用可能なクエリとしては，今回実験している中国語，ロシア語，ペルシャ語に翻訳されたものが用意されている．

4.2 実験設定

本研究においてのベースライン手法と実験の設定について述べていく．私たちは HC4 において行う実験のベースラインの 1 つとして，単語一致の手法である BM25 を用いた．BM25 ではクエリと文書で言語が異なっていると単語が一致せず検索ができないため，検索する対象の文書の言語に機械翻訳したクエリを利用した．本研究において，BM25 や Zero-Shot のように翻訳を利用する場合，検索の精度が翻訳の精度に依存せず，平等に評価できるように，全て Sockeye [6] とした．BM25 による検索では CLIR の実験を行うためのフレームワークである Patapsco [4] を利用した．neuclir1 におけるベースラインは，TREC 2022 NeuCLIR においてベースラインとして提供された結果を使用した．この結果は，文書を翻訳することでクエリと文書の言語を一致させ，BM25 で検索を行ったものである．HC4 のベースライン手法はクエリを翻訳することで文書の言語 (中国語，ロシア語，ペルシャ語) と一致させているが，neuclir1 では文書を翻訳することでクエリの言語 (英語) と一致させている．文書の翻訳はコストが高い一方で，精度が高いため強力でありベースラインとして使用されることがある．

HC4，neuclir1 においての実験の設定を以下で述べる．ColBERT-X の Zero-Shot や提案手法における学習では MS MARCO や neuMARCO などのデータセットを編集してクエリ，適合文書，非適合文書の組 (トリプル) を大量に用意にする．トリプルの構築方法について述べる．まず，クエリとその適合文書のペアを用意する．次にデータセットの文書集合において，そのクエリで BM25 を利用した検索を行い，ランキング上位 n 件の中からランダムに非適合文書を 1 つ抽出しそのトリプルの非適合文書としてこれら 3 つを 1 組とする．HC4 における実験では BM25 上位 1000 件，neuclir1 では上位 500 件から文書を選び，非適合文書として利用した．Zero-Shot では学習データが全て英語であり，提案手法では，クエリは全て英語で文書は英語，ロシア語，ペルシャ語，中国語のいずれかで表現されている．この時，各言語毎にデータの数は均等になるようにして，学習時に計算されるトリプルはランダムに並べる．学習時には，データが十分に用意できることからエポック数は 1 として，バッチサイズは 32，ステップ数 (パラメータを更新する回数) は HC4 の実験では 20 万回，neuclir1 では 10 万回とした．モデルには optimizer として AdamW を用いて，学習率は 3.0×10^{-6} とした．エンコーダとして XLM-RoBERTa-large を使用した．また，エンコーダには入力制限があり，一定以上の長さの文書は入力することができないという問題を解決するために文書を分割し，分割した部分毎にクエリとの適合度を計算した．分割をする際には，先頭にタイトルを付けた文書を 180 トークン毎に区切ってそれぞれの適合度を求め，求めた適合度の中で最も高い部分の適合度をその文書の適合度とした．ColBERT では自身の手法によって候補の文書を抽出し，その結果を並び替える「全文書検索」と，他の手法によって得られたランキングを並び替える「リランキング」が存在し，実験でも Zero-Shot と提案手法のどちらにおいても全文書検索とリランキングの検証を行った．各モデルの評価指標について述べる．HC4 では Recall@1000，nDCG@100，MAP を用いて，neuclir1 では Recall@1000，nDCG@20，MAP を用いて各モデルを評価した．また，提案手法によるモデルの統計的有意性

を評価するため，$p = 0.05$ の対応のある t 検定を実行した．

4.3 実験結果

表 3 に HC4 での実験結果を示す．提案手法と BM25 の間で統計的検定を行った結果，統計的に有意であると判断された場合 † で示している．提案手法と Zero-Shot との間で統計的検定を行った結果，統計的に有意であると判断された場合 $*$ で示している．提案手法の全文書検索の結果について述べる．中国語においては，Translate-Train が全ての指標において最も高い数値となっていて，提案手法は Translate-Train を上回ることができなかった．また，統計的検定の結果から中国語においての提案手法の全文書検索と BM25,Zero-Shot との結果の差は統計的に有意であったが他の手法との間では統計的に有意な差は認められなかった．ペルシャ語では提案手法による全文書検索が全ての指標において最も高い数値であり，提案手法と BM25 の結果の間で有意差が認められた．しかし，提案手法と他の手法の結果の間に有意差は認められなかった．ロシア語では，BM25 が Recall@1000 以外の指標において最も高い数値となった．しかし，全ての指標において提案手法の全文書検索と他の手法の結果の間に有意差がないことが分かった．

続いて，提案手法のリランキングの結果について述べる．中国語とペルシャ語ではリランキングを行った結果が BM25 の結果を改善しているが，ロシア語においてはリランキングした結果が BM25 の結果よりも低い数値を示している．また，統計的検定の結果からリランキングによる結果と BM25 による結果の間に有意差が認められたのは中国語のみであった．

実験結果から考えられることを述べる．中国語では，BM25 と Zero-Shot の 2 つの手法と Translate-Train と提案手法の 2 つの手法で結果に大きな差がある．これは BM25 と Zero-Shot の結果が低いことによるものだと考えられる．実際に，これらの手法の本実験での結果は HC4 や ColBERT-X の論文において示されている実験結果の値を再現できていない．これは BM25 や Zero-Shot がどちらも翻訳を利用していることから，翻訳が失敗しているために検索が上手くいっていないからであると考えられる．ペルシャ語では，BM25 を統計的に有意に上回っているため単語一致の手法と比較すると提案手法は有効であるといえる．Zero-Shot や Translate-Train と比較すると，これらの手法の結果を上回ってはいるが，統計的有意差が認められないため提案手法がこれらの手法と比較して有効であるとはいえない．ロシア語では BM25 や他の手法と統計的有意差はないものの結果を下回っているため，ロシア語では提案手法が有効に作用しているとはいえない．

提案手法の有効性についてより詳細に調査するために，別のテストコレクションでの評価結果を表 4 に示す．我々は，TREC2022 で開催された CLIR に取り組むタスクである NeuCLIR に参加した．NeuCLIR ではテストコレクションとして neuclir1 が与えられ，参加者は各自の手法による結果を提出する．そこで，neuclir1 に対して本研究の提案手法を適用し，その結果を提出した．表 4 はその評価結果である．neuclir1 ではベースラインとして，文書を翻訳することでクエリと文書の言語を一致させて BM25 で検索した結果が与えられたため，その結果を利用した．提案手法と BM25 の結果で統計的検定を行い，2 つの結果に有意差があると判断された場合，† で示している．

実験結果について述べる．提案手法による全文書検索では，ロシア語においてベースラインである BM25 の結果を統計的に有意に上回っている．また，全ての言語で，提案手法でベースライン (BM25) による結果をリランキングする手法が最も高い値を示した．特にロシア語では，Recall@1000 以外の指標において提案手法とベースライン間の差が有意であると判断された．また，中国語でも MAP においてベースラインの結果を統計的に有意に上回った．実験結果から考えられることを述べる．これらの結果は HC4 での結果とは異なる結果であることから，1 つのデータセットで結果からある言語について，必ずしも全ての文書に有効であるとは言えないことがわかる．Recall@1000 は上位 1000 件の結果に適合文書が多く含まれているほど高く，nDCG@20 や MAP は適合文書をより高い順位としているほど高くなる．提案手法は Recall@1000 が低いものの，他の指標が高くなっている．このことから，提案手法のモデルが適合度が高い文書を並び替えることに効果的であるため，リランキングに利用すると有効に作用していると考えられる．HC4 と neuclir1 に対する実験結果から，以下のような知見が得られた．

（1） 中国語では，提案手法は HC4 において BM25 を統計的に有意に上回ったが，言語に特化するような手法を上回ることはできなかった．neuclir1 では提案手法をリランキングに使用した場合に有効に作用する．

（2） ペルシャ語では，BM25 を統計的に有意に上回り，提案手法は HC4 において最も良い結果であった．neuclir1 ではリランキングに使用した場合にベースラインを上回ったが，有意差は認められなかった．

（3） ロシア語では，提案手法は HC4 において BM25 の結果を下回ったが，neuclir1 では全文書検索とリランキング共にベースラインの結果を統計的に有意に上回り，リランキングに使用した場合により有効に作用する．

（4） 提案手法の有効性は言語やデータセットによって異なり，リランキングを利用することが有効に作用する場合もある．

4.4 まとめ

本論文では，各言語のために利用していた各言語用の学習データを，まとめて 1 つのモデルのための学習データとして利用することで，単一のモデルによって各言語毎の言語横断情報検索を行うための学習方法を提案した．この方法では，他の言語の学習データを用いても目的としている言語でのタスクの結果が向上するという事前学習多言語モデルの特徴から，各言語毎のデータを 1 つの学習データとしてまとめた際にも各言語での検索の結果が向上すると考えた．実験では，CLIR のためのテストコレクションである HC4 と neuclir1 に対して提案手法を適用した．HC4 での実験から提案手法は中国語において BM25 を統計的に有意に上回ったが Translate-Train の結果を上回ることはできなかった．ペルシャ語では，他の手法を上

表 3　テストコレクション HC4 における実験結果

	クエリの言語	文書の言語	Recall@1000	nDCG@100	MAP
中国語					
BM25	中国語に翻訳	中国語	0.445	0.218	0.142
Zero-Shot	中国語に翻訳	中国語	0.575	0.308	0.185
Translate-Train	英語	中国語	**0.835**†	**0.552**†	**0.372**†
提案手法 (リランキング)	英語	中国語	0.448	0.327†	0.217†
提案手法	英語	中国語	0.813†	0.539†	0.361†
ペルシャ語					
BM25	ペルシャ語に翻訳	ペルシャ語	0.759	0.354	0.227
Zero-Shot	ペルシャ語に翻訳	ペルシャ語	0.808	0.445	0.270
Translate-Train	英語	ペルシャ語	0.774	0.396	0.232
提案手法 (リランキング)	英語	ペルシャ語	0.759	0.445	0.297
提案手法	英語	ペルシャ語	**0.842**	**0.475**†	**0.300**†
ロシア語					
BM25	ロシア語に翻訳	ロシア語	0.710	**0.347**	**0.223**
Zero-Shot	ロシア語に翻訳	ロシア語	0.672	0.336	0.217
Translate-Train	英語	ロシア語	**0.711**	0.340	0.220
提案手法 (リランキング)	英語	ロシア語	0.710	0.332	0.218
提案手法	英語	ロシア語	0.651	0.315	0.218

表 4　テストコレクション neuclir1 における実験結果

	クエリの言語	文書の言語	Recall@1000	nDCG@20	MAP
中国語					
BM25	英語	英語に翻訳	**0.781**†	0.340	0.264
提案手法 (リランキング)	英語	中国語	0.781	**0.396**	**0.286**†
提案手法	英語	中国語	0.563	0.364	0.222
ペルシャ語					
BM25	英語	英語に翻訳	**0.829**†	0.355	0.253
提案手法 (リランキング)	英語	ペルシャ語	0.829	**0.415**	**0.285**
提案手法	英語	ペルシャ語	0.591	0.330	0.200
ロシア語					
BM25	英語	英語に翻訳	**0.774**†	0.292	0.216
提案手法 (リランキング)	英語	ロシア語	0.774	**0.450**†	**0.321**†
提案手法	英語	ロシア語	0.601	0.366†	0.226†

回り特に BM25 を統計的に有意に上回った．ロシア語においては言語毎にモデルを用意する手法の結果を下回った．また，neuclir1 に対する実験ではロシア語において提案手法の結果がベースライン手法の結果を上回り，ペルシャ語では提案手法の結果がベースライン手法の結果を下回った．これは，HC4 での実験結果とは異なる結果であった．さらに HC4 と異なる点として，neuclir1 ではベースライン手法 (BM25) によるランキングを提案手法によってリランキングした結果が，全ての言語において最も良い結果となった．これらのことから，1 つのデータセットの結果からある言語について，必ずしも全ての文書に有効であるとは言えないことがわかった．また，リランキングの手法を選択することが効果的になる場合があることも明らかとなった．今後の課題としては，なぜ同じドメインのデータセットを使用しているにも関わらず 2 つの間で結果に大きな違いが出るのか，提案手法が有効でない場合にはどのようにすれば結果を改善できるか，今回実験した言語以外の言語での結果はどのようになるかなどが挙げられる．

謝辞　本研究は JSPS 科研費 22H03905，21H03554 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] Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34:7547–7560, 2021.

[2] Martin Braschler. Clef 2003–overview of results. In *Workshop of the cross-language evaluation forum for european languages*, pages 44–63. Springer, 2003.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal,

Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[4] Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. Patapasco: a python framework for cross-language information retrieval experiments. In *European Conference on Information Retrieval*, pages 276–280. Springer, 2022.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. The sockeye 2 neural machine translation toolkit at amta 2020. *arXiv preprint arXiv:2008.04885*, 2020.

[7] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[8] 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. *arXiv preprint arXiv:2004.04906*, 2020.

[9] 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*, pages 39–48, 2020.

[10] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[11] Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. Hc4: a new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer, 2022.

[12] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer, 2022.

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*, 2016.

[14] Tetsuya Sakai, Douglas W Oard, and Noriko Kando. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Nature, 2021.

[15] Peng Shi, He Bai, and Jimmy Lin. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, 2020.

[16] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. Cross-lingual training with dense retrieval for document retrieval. *arXiv preprint arXiv:2109.01628*, 2021.

[17] Ellen M Voorhees and Donna K Harman. *TREC: Experiment and evaluation in information retrieval*. MIT Press, 2005.

[18] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

[19] Eugene Yang, Suraj Nair, Ramraj Chandradevan, Rebecca Iglesias-Flores, and Douglas W Oard. C3: Continued pretraining with contrastive weak supervision for cross language ad-hoc retrieval. *arXiv preprint arXiv:2204.11989*, 2022.

[20] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3490–3496, 2019.