単一密検索モデルに基づく複数言語横断情報検索

阿部　健也

ユーザが入力するクエリから文書を検索するアドホック検索においてクエリ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) 」と呼ぶ．近年，密検索の CLIR への適用が進んでいる．CLIR では従来，クエリか文書の言語を翻訳してから検索を行う手法が主流であったが，この手法は翻訳に間違いがあると検索が失敗してしまう．しかし，密検索には単語の意味まで考慮できるという特徴があり，検索時に翻訳を必要としないため翻訳を利用する手法よりも有効であると考えられる．実際に，ColBERT-X では，密検索によって CLIR を行う手法を提案し，クエリを翻訳してから行う BM25 の検索よりも高い精度が得られた．ColBERT-X では言語毎にモデルを用意し，各モデルを言語毎に学習を行う手法が有効であった．複数のモデルを用意する手法は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．本研究では，単一のモデルで複数の CLIR をより良い精度で行える手法について考えた．密検索による CLIR では，事前学習多言語モデルというモデルを学習させて密検索モデルを生成する．事前学習多言語モデルは目的の言語以外の学習データを利用しても目的の言語のタスクの精度が向上することがわかっている．このことから，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しても検索の精度が向上すると考えた．我々は，複数の CLIR を行う事前学習多言語モデルの学習の手法として，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する手法を提案した．実験ではデータセットとして HC4，neuclir1 を利用し，それぞれのデータセットでベースライン手法と提案手法の結果の比較を行った．HC4 では，追加の実験として ColBERT-X で有効であった言語毎にモデルを用意して学習する Translate-Train との比較も行った．HC4 での実験の結果，ペルシャ語では提案手法がベースライン手法および Translate-Train を上回った．中国語ではベースライン手法を上回ることができたものの Translate-Train の結果を上回ることができず，ロシア語では実験した手法の中で最も低い結果となった．neuclir1 に対する実験ではロシア語において 2 つの指標で提案手法の結果がベースラインの結果を上回ったものの，ペルシャ語と中国語では提案手法の結果がベースラインの結果を下回った．また，neuclir1 ではベースライン手法のランキングを提案手法によってリランキングする手法が全言語で最も良い結果であった．これらの結果は HC4 での結果とは全く異なる結果であることから，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る．HC4，neuclir1 に対する実験結果から提案手法はデータセットや言語によって効果的に作用すること，リランキングを利用することが有効に作用する場合があることが明らかとなった．

（指導教員　加藤　誠）