質問生成と機械読解に基づく
情報検索アルゴリズム

薄羽　皐太

情報検索の中心的なタスクであるアドホック検索タスクは，与えられたクエリに対して，適合度が高い順に文書を並び替える問題である．これとは異なるタスクとして機械読解タスクがあり，これは質問と文章が与えられた時に，文章からその質問に対する答えを抽出するタスクである．アドホック検索タスクは長年にわたって取り組まれてきた問題であるのに対して，機械読解タスクは近年の大規模なデータセットの登場と機械読解モデルの性能の改善を背景に，特に注目を集めている．これらの問題はそれぞれ別の問題として取り組まれ発展してきたが，ある文字列に合致する内容を取得するという面では目的が一致しており，これまでに提案されてきたモデルには多くの共通点が存在する．アドホック検索では，クエリに合致するような内容の文書を文書群から取得することを目的とし，機械読解タスクでは質問に合致するような回答を文章中から取得することを目的としている．この 2 つの異なる問題の共通点に着目し，本論文では機械読解モデルを用いることによってアドホック検索タスクを解く方法を提案する．アドホック検索タスクを機械読解モデルで解くために，検索クエリからその背後にある質問を生成し，その質問に対する答えを含むかどうかを機械読解モデルによって判定することで，文書の適合性を推定する．機械読解モデルによってアドホック検索の問題が十分な精度で達成できるのであれば，機械読解モデルの発展がそのままアドホック検索の問題の貢献へとつながり，より効率的な技術発展が期待できると考えた．機械読解モデルとして事前学習済みの BERT を機械読解タスクのデータセットである SQuAD2.0 でファインチューニングしたモデルを用いる．機械読解モデルに質問と文書を入力し，出力として得られる文章中のある位置から答えが始まる確率の最大値を適合度として利用した．検索クエリから質問を生成する手法は，文字列から文字列への変換であるため，機械翻訳タスクとして捉え，機械翻訳モデルを用いた．機械翻訳モデルの学習には，質問からクエリを生成することで，クエリと質問がペアになるようなデータセットを構築し，学習した．実験では提案手法のアドホック検索の性能の評価を行なった．データセットにはアドホック検索タスクである NTCIR WWW-2・WWW-3 の English タスクのデータセットを用い，アドホック検索についての性能をベースラインと提案手法を比較した．さらに，質問生成の手法をいくつか検討し，質問生成の改善がどのように機械読解モデルによる文書のランク付けに影響するかを検証した．また，機械読解モデルによる適合性推定の際に得られる答えについて分析を行なった．その結果，BM25 との比較において，機械読解モデルによる文書のランク付けを部分的に行うことで，アドホック検索の性能が改善することが判明した．

（指導教員　　加藤　誠）