密検索モデルのフレーズクエリ解釈能力の評価

土戸　翔太

検索品質は、ユーザーが求める情報に迅速かつ正確にたどり着けるかを示す重要な指標であり、多くの EC サイト等のや検索サービスでは、検索品質がサイトの収益やユーザー満足度に深く関わると考えられている。近年、BERT などの学習済み言語モデルを用いた密検索（Dense Retrieval）が注目を集めており、クエリと文書をベクトル空間へ投影することで、曖昧なクエリや同義表現にも柔軟に対応できる利点をもつ。しかし、単語単位のインデックスを構築しない性質上、従来のキーワード検索で容易に実現していた演算子機能（フレーズ検索や NOT 演算子など）がどの程度正確に扱えるかは、依然として十分な検証が行われていない。結果として、演算子を正しく解釈できない場合には、ユーザーが特定の単語列を連続して検索したり、特定語を除外したりといったニーズを満たせず、検索体験の質が大きく損なわれるおそれがある。本研究では、特にフレーズ演算子「“”」に着目し、密検索モデルが複数単語の連続一致をどの程度厳密に捉えられるかを明らかにし、部分的な改善策を提案することを目的とする。具体的には、BM25 を用いて、キーワード検索を実装し、密検索と比較した。部分的な改善策として、MSMARCO Passage データセットに含まれるクエリと文書のペアを再分類し、フレーズ演算子を考慮した学習データを作成した。その際、クエリ内に含まれる複合名詞が文書中に存在するか否かを厳密に判定する「厳密再分類型データセット」と、既存の正例を保持しつつ追加で複合名詞が確認された文書を正例に加える「既存正例継承型データセット」の二種類を構築した。これらを用いて Dense Passage Retrieval（DPR）モデルを学習させることで、密検索モデルのフレーズクエリ解釈能力を改善することができるのかを明らかにした。さらに、再分類データセットによる単独学習だけでなく、ベースラインモデルに対する追加学習（fine-tuning）の効果も併せて評価した。実験の結果、キーワード検索をベースとする BM25（Elasticsearch）で“”演算子を実装した場合が最も高い検索精度を示し、フレーズ検索を厳密に扱う点では疎検索の方が優れていることが明らかになった。一方、いずれの再分類データセットでも NDCG@10 が既存の MSMARCO Passage データセットを学習させたベースラインを上回り、学習データの再定義によって密検索モデルがフレーズ演算子の解釈能力を向上させることが示された。また、MSMARCO Passage データセットを学習済みモデルに対して再分類データセットを用いた追加学習を行ったところ、さらに検索精度が向上する傾向があり、クエリ演算子の解釈精度向上には学習の形態やデータセットの設計が大きく影響することが示唆された。

（指導教員　　加藤　誠）