クエリ自動補完のためのコーパスからのクエリログ生成

染谷　瑛進

情報検索システムには，クエリ自動補完 (QAC) という機能が採用されている．QAC は，ユーザが入力途中の不完全なクエリから最終的に入力しようとしている完全なクエリを予測し，その候補を提示するタスクのことを言う．QAC には，ユーザの検索をサポートする効果があり，現在の検索システムにおいては必須の機能である．QAC のアプローチには様々なものがあるが，その多くは，蓄積された過去のクエリログを用いる手法である．しかし，十分なクエリログを入手できない場合，このような手法を使用することはできない．ただし，そのような小規模の検索システムであっても，どのようなトピックのクエリが使用されやすいかについては，ある程度予測できる場合も多い．そこで，本論文では，検索サービスにおいて，ユーザが実際に使用しうるクエリの集合を生成する手法の提案を行う．具体的には，検索対象となる文書集合を入力とし，それを何らかのクエリ生成システムに入れることによって，クエリの集合を生成する．ただし，本論文では，目標とするクエリ集合のトピック分布は入手可能であるという前提のもと，クエリログを代替するクエリ集合の生成を行う．クエリの生成には，文書からクエリを生成する手法であり，高いランキング性能と応答速度を兼ね備える docTTTTTquery を使用する．まず，入手可能な他のクエリ集合とそれとペアになる文書集合を用いて，docTTTTTquery においてクエリを生成する事前学習モデルとして用いられる T5 言語モデルに対し，文書からのクエリ生成パターンを学習させる．次に，検索サービス内に含まれる文書集合について，得られたトピック分布と入力として与える文書集合のトピック分布が同様の分布になるように調整を加えた上で，文書を 1 つずつ取り出し，それを学習済みの T5 言語モデルに入力することにより，クエリ集合を生成する．実験では，ORCAS と MS MARCO の 2 つのデータセットのペアを使用し，MS MARCO 中の文書から様々な条件のもと単語を抽出することで擬似クエリログを生成し，これをベースラインとした．また，提案手法として，AOL クエリログと AOL 中のクエリによりクリックされた Web ページの集合をペアとして与え，クエリ生成パターンを学習させた T5 言語モデルに対し，MS MARCO 中の各文書を入力として与えることで擬似クエリログを生成し，ベースラインとの比較を行った．実験により，提案手法は，ベースラインと比べて有効性がないことが示された．しかし，入力文書集合のトピック分布を調整することは有効である可能性がある．そこで，提案手法と同様に，トピック分布を調整した MS MARCO の文書を入力とし，単語抽出によるクエリ集合の生成を行う実験を行った．その結果，この手法がベースラインと比較して有効であることが示された．また，予備実験において，抽出語数を 4 語としてクエリを生成した時，最も有効であることが示された．

（指導教員　　加藤　誠）