「思い出せない」既知アイテム検索における誤りを含むクエリの処理方法

冨澤 燦之丞

映画などのタイトルが「思い出せそうで思い出せない」現象を Tip of the Tongue (ToT)現象という．この現象は多くの人が経験するものであり，英語圏の大手掲示板 Reddit では 250 万人を超えるユーザーが ToT コミュニティに参加し，思い出せない映画や本，テレビ番組や音楽に関する質問を寄せている．これは Reddit コミュニティの中でも特に規模が大きく，多くの関心が寄せられている．また，ToT 状態は大きな心理的フラストレーションを引き起こすことも報告されており，効果的な解決策が求められている．ToT 状態においてユーザーの記述したクエリ（ToT クエリ）は，記憶の曖昧さに起因する不正確な情報を含むため，従来の検索システムでは十分な対応が困難である．例えば，代表的な検索モデルである BM25 は，クエリと文書中の単語の一致に基づいて検索を行うが，ToT クエリには誤った単語が含まれている可能性が高く，適切な検索結果を得ることが難しい．この問題に対し，既存研究では大規模言語モデル（LLM）を用いてクエリから誤情報を取り除くアプローチが提案されている．しかし，この手法では誤情報の除去と同時に，検索に有効な手がかりまでも失われてしまうという問題がある．また，他にも Virtual Adversarial Training という手法を用いて，曖昧なクエリに対処する手法も提案されている．この手法では，正確な情報を含むクエリと，意図的にノイズとなる情報を加えたクエリの両方で同じ出力が得られるようにモデルを学習させる．これにより，誤情報を含むクエリでも適切な検索が可能になるが，モデルが過度にノイズに対して寛容になり，本来区別すべき情報を同一視してしまうという課題がある．そこで本研究では，ToT クエリ中の不確実な情報を削除せずに活用する手法を提案する．提案手法では，LLM を用いてクエリ中の各単語に対し書き手がどの程度確信を持っているかを数値化し，その値を利用して ColBERT を基盤とする検索モデル内でスコアの重み付けを行う．この手法では，前後の文脈や文章表現を踏まえて確信度を推定することで，曖昧な情報や誤情報として扱うべき部分を明確に区別することができる．また，スコア計算でこれらの情報を活用することで，手がかりを失うことなく検索精度の向上を図ることができる．ColBERT 内での重み付けにおいては，それぞれ異なる設計の関数 3 つを作成して実験を行った．実験では TREC ToT のデータセットを用いて，従来の手法（ColBERT，BM25，DPR，LLM トリミング+ColBERT）と MRR や Recall などの指標において精度の比較を行った．その結果，重み付け関数の設計によって異なる評価指標で改善が見られ，確信度の反映方法によって目的に応じた検索性能の改善が可能であることが示された．

（指導教員　加藤 誠）