逆学習を用いた検索用学習データセットの蒸留

加藤　匡一郎

近年，大規模言語モデル（LLM）の進展により，検索エンジンの性能は大きく向上している．LLM は自然言語処理の分野で革新をもたらし，特に情報検索において重要な役割を果たしている．従来の検索エンジンはキーワードマッチングやルールベースのアプローチに依存していたが，LLM は文脈を理解し，意味的な関連性に基づいて情報を抽出する能力を持つ。これにより、ユーザーが検索クエリを入力すると関連性の高い文書を迅速かつ正確に見つけることが可能となった．例えば，「ストレス 不眠症」というクエリに対して，LLM は文書を密ベクトルで表現し，内積計算を通じてスコアを算出して関連度に基づいてストレスや，不眠症に関連した文書を順位付けする．このプロセスは単なるキーワード一致を超え，文脈やテーマの関連性を考慮することでより精度の高い検索結果を提供する．しかし，LLM を用いた検索モデルの学習には課題も存在する．その一つが，膨大な学習データを用いる際の効率性の問題である。すべてのデータがモデルの性能向上に等しく貢献するわけではなく，影響の小さいデータが学習効率を損ねる場合がある．この結果、計算コストが増大し、学習時間が長期化するなど、リソースの無駄遣いが発生する可能性がある．この問題を解決するためには，影響度の低いデータを除去するなど，効率的なデータ選定が求められる．このような背景の中で注目されている手法が「逆学習」である．逆学習は学習データの中から影響力の大きいデータを特定し，影響力の小さいデータを除去することで，モデルの学習効率を向上させるアプローチである．この手法では，データごとの貢献度を定量的に評価し，学習時間を短縮しながら性能を維持することを目指す．逆学習によって学習データの重要性を測定し，効率的なデータ選定を可能にする枠組みが提供される．本研究では，逆学習を利用して検索モデルの学習効率を向上させる手法を提案する．提案手法では，学習データごとの貢献度を解析し，効率的に選別するプロセスを体系化した．本研究の目的は，検索モデルの性能を維持しながら学習時間を短縮することで，より効率的なモデル構築を実現することである．具体的には，UnTrac-Inv という逆学習フレームワークを採用し，当初は MS MARCO データセットを用いた実験を計画していたが，最終的に SciFact を用いて実験を行い，性能評価には nDCG を使用した．本研究は，検索モデルの学習におけるデータ選定プロセスの重要性を明確にし，逆学習による効率化の可能性を示した．ただし，データ収集の不足や評価条件の不備といった課題が残されている．今後は，より多くのデータを収集し，条件設定を適切に行うことで，さらなる精度向上と一貫性のある結果の提供を目指す．

（指導教員　　加藤　誠）