市民ツイートの回答らしさを考慮した応答文選択手法の提案

中山　聖司

日常生活の中で生まれる疑問の中には同じ地域に住む人にしか答えることができないものがある．しかし，現状の情報サイトは地域性が弱い，情報元が限定される，新しい情報が集まらないなどの欠点がある．

そこで，本研究では，Twitter データを知識リソースとし，質問に対して回答となる市民ツイートを，関連度と回答らしさを考慮して選択する手法を提案する．質問とツイートの関連度は，良質な文書ベクトルの生成ができる Sentence BERT を用いて構築した，質問とツイートの文の分散表現間の cosine 類似度として計算する．回答らしさは，Yahoo! 知恵袋のベストアンサー約 10,000 件と類似したツイートを 50 件ずつ抽出し，抽出したツイートごとに各順位の逆数を加算した合計値として計算する．また，この際，回答らしさを考慮するために，回答文と質問文とを分類するタスクでファインチューニングした BERT により構築された文の分散表現間を用いて cosine 類似度を計算する．

実験では，MAP，MRR，P@1，P@5，nDCG@5 の評価尺度を用いて，提案手法と 4 種類の比較手法を評価する．比較手法は，(1) 関連度だけでランキングする手法，(2) 回答らしさを考慮するための BERT のファインチューニングを行わない手法，(3) Yahoo!知恵袋の回答データを保育園関連の回答データに変更する手法，(4) (2) と (3) の変更を同時に行う手法 の 4 種類である．

提案手法は，P@1，nDCG@5 において，t-検定（有意水準 5%，両側検定）で，関連度だけでランキングを作成する手法以外の 3 つの比較手法に対して有意に向上していることを確認した．一方で，MAP， MRR，P@5 については，有意な向上を確認できなかった．この理由としては，提案手法による回答らしさの判別基準は，ランキングならびに最上位候補に反映されているものの，回答らしさが明確な要素がない場合には，関連度のみで回答候補を獲得する比較手法などと，集合を評価するこれらの指標において差が出にくいことが影響したと考える．

（指導教員　関洋平）