Twitter 特有のコミュニケーション表現の抽出

風間 千明

Twitter では，「なう」や「わず」といったような，特徴的で簡潔な表現が文末に多く見られ，これらの表現はコミュニケーションにおいて重要な役割を果たす．本研究では，このような表現を「コミュニケーション表現」とし，Twitter 特有のコミュニケーション表現を自動抽出するための手法を提案した．

自動抽出をする上では，コミュニケーション表現が文末に出現しやすいことと，文末記号の違いによってその前に出現する文字列にも違いが出ることを利用した．まず，文末記号に応じて Twitter のツイートに高頻度で出現する多様な文末ひらがな 2-gram（「なう」，「あり」など）を抽出する．次に，Twitter のツイートに高頻度で出現する文末ひらがな 2-gram と他の文書ジャンルのテキストに高頻度で出現する文末ひらがな 2-gram とを比較することで，Twitter 特有の文末ひらがな 2-gram を選択する．そして，Twitter 特有の文末ひらがな 2-gram を含み，その前に接続するひらがな n-gram 表現をコミュニケーション表現候補として抽出する．さらに，コミュニケーション表現候補からノイズを減少させるために，Twitter 用語集に収録されている語を利用して頻度の下限を設け，低頻度の文末ひらがな n-gram を除去し，n-gram 確率を利用してコミュニケーション表現候補を絞り込む．

実験では，文末ひらがな n-gram と文末記号との組み合わせの頻出要素について，コミュニケーション表現が含まれるかを調査した．調査の結果，リプライツイートを利用するほうが，より多くのコミュニケーション表現を抽出できること，文末記号の違いを用いることで，多様なコミュニケーション表現を抽出できることがわかった．

そして，文末ひらがな n-gram(n=2∼5) についてコミュニケーション表現候補の絞り込みを行うために，Twitter 用語集に収録されているひらがな n-gram(n=2∼5) と n-gram 確率を利用し，絞り込みのための閾値の設定について検討した．サンプルの出現傾向から再現率・精度・F 値を計算した．精度については，サンプルの含まれる割合が大きくなるにつれて少しずつ低下してしまうが，サンプルの含まれる割合が 5 割から 6 割の間では精度の落ち方が緩やかであることがわかった．また再現率の観点からは，多くのサンプルが含まれることが好ましい．以上のことから，サンプルの含まれる割合が 6 割の場合の結果を，閾値として利用することが妥当であると判断した．

今後の課題としては，n=6∼10 の場合の文末ひらがな n-gram のコミュニケーション表現候補の絞り込みを同様に行うことが挙げられる．

（指導教員　関 洋平）