トピックモデルによる単語の分散表現手法に関する研究

野沢　健人

単語の意味を理解することは，より大きな言語構造を理解する上で重要な要素となる．そのため，単語の文法的・意味論的な理解に関する研究は重要な研究課題として位置づけられる．近年，分布仮説に基づいた単語の分散表現の獲得手法に関する研究は，応用を含め盛んに研究されている．しかし，獲得した分散表現において，単語ベクトルの空間と意味の対応関係について解釈できない課題がある．

本手法では，トピックモデルを用いた単語の分散表現獲得手法を提案する．分布仮説に基づき，1 つの単語タイプをその周囲で出現する多重集合で表し，トピックモデルを適用することで単語を多項分布で表現する．この多項分布をベクトルとみなすことで単語の分散表現を獲得する．ベクトル演算の他に提案手法で獲得した分散表現は，Jensen-Shannon ダイバージェンスなどの確率分布に対する演算を用いた単語間の類似度計算が可能になる．

既存手法との比較実験として word similarity と analogy のタスクによる評価を行い，実験結果から提案手法はベクトルの次元数が高いほど，それぞれのタスクの評価値が高くなるモデルであり，既存手法は，タスクに応じて適切なベクトルの次元数は異なるモデルであることが明らかになった．また，実験結果から提案手法はそれぞれのタスクにおいて，既存手法を上回る評価値を得ることはできなかった．原因の 1 つとして多項分布をベクトルとみなしても，既存手法のようにベクトルに単語の意味が反映されていないためであることが考えられる．このため，単語を確率分布で表現した場合に，ベクトルの加算減算に対応する演算の定義が今後の課題である．

提案手法では，トピックモデルの学習結果からトピック分布と単語分布を得ることができるため，分散表現で対応付けた単語におけるトピックの確率値を用いてどのトピックが高く割り当てられているかを比較できる．トピックモデルを用いているために，トピックがどのような単語によって特徴付けられているかを確率値とともに獲得できる．このため，それぞれの単語がどの単語によって特徴付けられているかをトピックごとに求められることから，従来手法では困難であった次元の解釈が可能なモデルとして位置づけられる．次元の解釈の例として，*bow* という英単語に対して，船首の意味とお辞儀をするという動詞の品詞の異なる 2 つの意味を別々のトピックとして獲得できたことから，本提案手法において単語の意味をトピックとみなすことでベクトルの次元が解釈可能であると結論づける．

（指導教員　若林　啓）