平成29年度 情報工学コース卒業研究報告要旨

旧金森 研究室	氏　　名	松山　仁
卒業研究題目	負例サンプリングによるSkip-gramの学習方法に関する性能評価	

近年、自然言語処理の分野で神経回路（ニューラルネットワーク）を用いる方法が多く提案されている。これらのモデルでは、入力として近年自然言語処理の分野で活用されている手法の一つである分散表現がよく用いられる。分散表現とは単語を200次元ほどのベクトルにより表現する手法であり、単語をニューラルネットワークの入力とする際に各単語の分散表現によるベクトル値を用いることができる。この分散表現を取得する方法のひとつとして、ニューラル言語モデルの派生である対数双線形モデルが提案されている。その中で、Thomas Mikolovらが提案している新しい対数双線形モデルのひとつが、Continious Bag-of-Words モデル、およびSkip-gramモデルである。この二つのモデルは合わせてword2vecと呼ばれており、分散表現の取得法として現在多く利用されている。

Skip-gramモデルを利用する上での問題点の一つは計算のコストの大きさであった。そこでMikolovらの研究により、その学習の高速化を実現する手法が提案されている。それが階層的ソフトマックスによる高速化、および負例サンプリング学習による高速化である。前者は単語群をいくつかのまとまりに分割して階層的に学習を行う手法であり、後者はすべての単語群に対して重みの更新を行う代わりにいくつかの偽の入力を用いて学習の近似を行う手法である。この研究に対して、より実用的なアルゴリズムとして先述の負例サンプリングによる学習をオンライン学習のアルゴリズムに拡張した理論的な手法が鍛治らによって提案されている。

本研究では負例サンプリングによるSkip-gramモデルおよびオンライン学習のアルゴリズムを用いた負例サンプリングによるSkip-gramモデルについて自分の実行環境の下で実際に実験を行い、負例サンプリングによるSkip-gramモデルの特性について考察した。また、実験や考察の過程で考えられた問題点について指摘し、考察を行った。

まずはじめに負例サンプリングによるSkip-gramモデルについて実験を行った。分散表現取得のための訓練データとしてインターネットで公開された様々なニュースの文章を抽出してまとめた'text8'を用いて、負例サンプリングによるSkip-gramモデルとオンライン学習によるモデルとの比較を行った。比較の手法として、単語間コサイン類似度とベンチマークの単語間相関度とのスピアマン順位相関係数を用いた。実験の結果、両モデルの値はともに0.6付近の類似した結果を得られた。精度が先行研究からの予想より下回ったのは、負例サンプリングの抽出数などのベースライン設定を含む実行環境の差異によるものと考えられる。

次に学習方法の改善について検討を行った。負例サンプリングによるSkip-gramモデルのアルゴリズムには、出現頻度の高さに応じて学習を行わない単語を選択するサブサンプリングとよばれる構造が存在する。このサブサンプリングについて、単語の出現頻度に関連させて単語を選択するのではなく、一定の出現頻度を超える単語をまとめて学習させないようにするアルゴリズムとし、その場合に精度が向上するのではないかという考察を行った。しかし確率の取り方を変更しながら実際にベンチマークの存在する英語のコーパスに関して実験を行ったところ、目立った精度の向上はみられなかった。

サブサンプリングにおける学習しない単語の適切な選択方法は、訓練データによって異なることが多い。これを判断し、より適切に選別することが今後の課題である。