令和3年度 コンピュータ科学科卒業論文要旨

外山 研究室	氏　　名	加 藤　辰 弥
卒 業 論 文 題 目	文の意味的類似度評価におけるグラフ構造の利用	

文の類似度を求めることは, 翻訳の性能評価のほか, 迷惑メールやスミッシングの検出などに有用である．文の類似度を求める手法としては, 編集距離や共通する単語数, さらには構文情報を用いた手法などが利用されてきたが, いずれも表層的な類似度を計算するだけで, 意味的な類似度を考慮していない．そこで, 文の意味的な類似度を評価する STS タスクが提案されている. STS タスクでは, 文 “Kids in red shirts are playing in the leaves” と文 “Children in red shirts are playing in the leaves” は意味的類似度が高く, 文 “A woman is riding a horse” と文 “A man is opening a small package that contains headphones” は意味的類似度が低いと評価する.

ニューラルネットワークの発展に伴い, STS タスクでは LSTM や RNN などの自然言語処理モデルを用いた手法が提案された．LSTM を用いた手法では, Elvys らがピアソンの相関係数 0.8549 という結果を示している．LSTM や RNN は文の意味情報だけを利用し, 文の依存関係のような構文情報を利用していない．意味情報に加えて構文情報も扱うことにより, 文を表す特徴量を増やすことが期待できる．構文情報はグラフで表現できる.

そこで, 構文情報からのアプローチとして, グラフ構造と, それを処理するためのグラフニューラルネットワーク (GNN) の利用を検討した．代表的な GNN として, Graph Convolutional Network (GCN) や Graph Attention Network (GAT), GraphSAGE といった手法がある. また, グラフ編集距離を近似的に求める SimGNN や, GCN を用いたグラフプーリング手法である Self-Attention Graph Pooling (SAGPool) も提案されている.

本研究の目的は, STS タスクにおいて, グラフ構造を用いたアプローチの有効性を明らかにすることである.

実験は, 文からのグラフの作成, ノードの畳み込み, グラフを表現する埋め込み生成, グラフ間の類似度の評価という流れで行い, それぞれの過程で 2 通りずつの方法を採用し, 計 16 通りの方法を試した. 実験の評価にはピアソンの相関係数を利用した. データセットには, 短文データセットである Sentences Involving Compositional Knowledge (SICK) と混在長データセットである STS Benchmark (STS-b) の二つのデータセットを用いた.

実験の結果, 16 通りの方法のうち, グラフの作成には単語の依存関係に基づく構築法を, ノードの畳み込みには GCN を, グラフを表現する埋め込みの生成にはアテンションモジュールを, グラフ間の類似度評価にはニューラルテンソルネットワークをそれぞれ利用した方法において最も優れた結果が得られた．この方法では SICK で, ピアソンの相関係数 0.848 ± 0.014 という結果が得られた. これは Elvys らの結果と同等であり, このことは STS タスクにおけるグラフ構造の利用の有効性を示唆している．しかし, STS-b では十分な結果を得られなかった．この理由として 次の二つが考えられる．第一に, 作成したグラフには構文情報は含まれていたが, 意味情報が十分に含まれていなかったことが挙げられる. これは, 単語の埋め込みの生成に文脈を考慮しないモデルを利用したことに起因していると推測される. 第二に, グラフサイズに差が生まれることが挙げられる. これは, SICK は短文データセットであるのに対し STS-b は混在長データセットであることに起因していると推測される.

将来の展望としては, 単語の埋め込み生成に文脈を考慮したモデルを用いることで, 今回のモデルの欠点を補うことが期待できる.