ニューラルトピックモデルに基づく解釈性の高い将来予測モデル

中村　礼音

大規模文書集合から，有益な情報を得るためのツールとしてトピックモデルがある．トピックモデルを用いることにより，人手を介在させることなく，大規模文書集合から話題になっているトピックを抽出することができる．また，文書ごとのトピックの配分率を知ることができる．ベイズ推定（変分ベイズ推定やギブスサンプリング）に基づく潜在ディリクレ配分法（LDA）や変分オートエンコーダ（VAE）推論に基づくニューラルトピックモデルは（NTM）はトピモデルの代表的な手法であり，大規模文書の分類，クラスタリング，検索などさまざまな応用タスクに利用されてきた．

一方，既存のトピックモデルでは未観測の将来の潜在的トピックを予測することができない．トピックモデルの活用領域の広さを考えると，時系列文書を想定したトピックの将来予測は重要な問題であると考えられる．より具体的には，ネット記事のトピックや特許文書の技術トピックのトレンドがこれまでどのように変化してきたか，あるいは今後どのようなトピックが盛り上がっていくのか，などといった洞察を得たい場合が挙げられる．

そこで本研究では，時間情報をもつ文書集合から，「将来の潜在トピック」あるいは「将来のトピックを構成する文書量」を予測するためのモデルを提案する．以下，このモデルを「ニューラルトピック将来予測モデル」と呼ぶこととする．より良いニューラルトピック将来予測モデルを構築するために，いくつかのモデルを導入して実験的に評価を行う．その中の一つの手法は，次のようなものである．意味解析を行うためのツールとしてNTMを，予測を行うためのツールとしてLSTMを導入する．次にNTMとLSTMの損失関数をそれぞれ結合することによって同時にモデルの最適化を行う．またこの手法の変種として，同時に最適化を行わず順に最適化を行うというものも考えられる．

ここで，時間情報をもつ文書集合の意味解析を行う際は，それぞれの期間で整合性のあるトピックを生成する必要がある．そこで，あらかじめ訓練データ全体で学習した事前学習済みニューラルトピックモデルを構築し，事前学習済みNTMのパラメータ全体を各期間で学習するNTMの初期パラメータとして受け継ぐことで，異なる期間でも整合したトピックが生成されるように工夫する．

実験では，「1999年から2019年までの建築分野の特許文書」と「2019年6月から2022年２月までの新型コロナウイルスに関するYahooニュース記事」をデータセットとして使用し，提案手法の検証を行った．トピックモデルの性能評価指標として，モデルの汎化性能を図る指標であるPerplexityを使用した．既存手法であるLDAとのPerplexityスコアを比較することで，トピックモデルの有効性を比較した．また，より良いニューラルトピック将来予測モデルを発見するために，平均二乗誤差（MSE）を用いて予測の性能評価を行った．

（指導教員　　伊藤　寛祥）