様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ３ 日現在

機関番号：３２６８６
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１４４０

研究課題名（和文）トピックモデルにおけるRNNの利用の有効性に関する研究

研究課題名（英文）Research on the effectiveness of using RNN in topic models

研究代表者
正田　備也（MASADA, Tomonari）

立教大学・人工知能科学研究科・教授

研究者番号：６０４１３９２８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：トピックモデルは、膨大な文書集合の中で扱われている多様な話題を自動的に抽出する技術である。しかし、トピックモデルによる通常のテキスト分析は、各文書での単語の出現頻度は考慮するが、語順は考慮しない。そこで、文書を単語列としてRNN（回帰型ニューラルネットワーク）でモデリングすることにより、トピックモデルを拡張する。RNNとトピックモデルを組み合わせる研究はすでにあるため、それらとは異なる手法を提案する。結果、従来とは異なる手法で文書を単語列としてモデリングするトピックモデルを提案した。だが、単語列をMLP（多層パーセプトロン）でモデリングするに留まり、RNNを組み合わせるまで到らなかった。

研究成果の学術的意義や社会的意義
20年近くテキストマイニングに使われてきたトピックモデルと、新しい技術である深層学習とを、どのように組み合わせれば効果的なテキストマイニングが実現できるか。この問いに本研究は取り組んだ。成果としては中途の段階ではあるが、従来研究でこの組み合わせを実現するために使われている変分オートエンコーダとは異なるアイディアにもとづいて、トピックモデルと深層学習を組み合わせる可能性が確かに見えたのは重要な成果である。この方向で研究を続ければ、膨大な文書集合に潜む多様な話題を抽出するツールとしてのトピックモデルを、深層学習による言語データのモデリングと組み合わせることで、さらに強力にすることができるだろう。

研究成果の概要（英文）：Topic models, including LDA (latent Dirichlet allocation), can automatically extract semantically meaningful themes from a large corpus. However, text analysis using topic models often only considers word frequencies in a document and does not consider the way words are arranged. This work aims to improve topic models with RNN (recurrent neural network) for modeling word order. Several previous studies propose a method for combining RNN with topic models. Therefore, we have tried to propose a new method. As a result, we have proposed a new topic model using NNs (neural networks), where we perform no VAE (variational autoencoder) inference. We instead maximize the target given in the original LDA paper by training NNs in an amortized manner and obtaining posterior parameters as output of NNs. However, we currently only use MLP (multilayer perceptron) and thus have not achieved our goals yet. We now have a plan to replace MLP with RNN or other more recent NN architectures in near future.

研究分野：機械学習

キーワード：機械学習　テキストマイニング　トピックモデル　深層学習

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

深層学習がテキストマイニング分野にも影響力を強めてきた 2016 年末以降、トピックモデルに RNN (Recurrent Neural Network)を組み合わせる提案が複数現われた。潜在的ディリクレ配分法 LDA [Blei+ 03]をはじめとするトピックモデルは bag-of-words モデルであり、単語の並びの前後関係を考慮することはできなかった。しかし、RNN と組み合わせることでトピックモデルにおいて文書の構造を考慮する可能性が拓かれた。とは言え、bag-of-words モデルのシンプルさを捨てて語順を考慮することで、従来とはどのくらい違うことが、トピック語抽出（図 1）などの応用で実現できるのだろうか。このような問題意識から本研究は出発した。

図1．UCI機械学習リポジトリのNY Times データから抽出されたトピック語の例

２．研究の目的

(1) 研究開始時の研究目的

研究開始当時、トピックモデルに RNN を組み合わせた提案が少なくとも 3 つあった。TopicRNN [Dieng+ arXiv:1611.01702]、潜在的 LSTM 配分法[Zaheer+ ICML17]、SenGen [Nallapati+ arXiv:1708.00308]である。そこで、研究代表者が取り組んで来たさまざまな応用課題にこれらの手法を適用すると、今までとは違って何ができるようになるかの精査を研究目的とした。しかし、この研究目的は、自然言語処理分野の大きな変化により変更されることとなった。

(2) 変更後の研究目的

2017 年に提案された Transformer[Vaswani+ arXiv:1706.03762]が自然言語処理の研究動向を大きく変えてしまった。特に、LSTM や GRU などの RNN が単語列のモデリングに使われなくなっていった。その結果、RNN にこだわる意味が学術的に希薄となった。本研究の開始当時は、これほど大きな変化になると予想しなかったため、2018 年度の間は、まだ当初の研究目的に応じるかたちで、RNN の学習そのものに馴染むための短歌の自動生成や、トピックモデルを深層学習的な方法論で訓練する作業に馴染むための変分推論の改良などに取り組んでいた。

だが、2019 年度に研究目的を修正した。具体的には、RNN であれ Transformer であれ、文書中の各単語トークンについて別々に分散表現を出力するニューラルネットワークを利用するような、トピックモデル用の変分オートエンコーダ(VAE)を提案すること、である。これを新しい研究目的として設定した。

当時知られていた深層学習版トピックモデルであるニューラルトピックモデル[Srivastava+ arXiv:1703.01488]は、単語トークンのトピック割り当てを表す離散潜在変数を周辺化して消去する VAE を使っていた（図 2）。しかし、Transformer 系の言語モデリングが主流になっていくと、言語データを単語列として明示的にモデリングし、単語トークンごとの分散表現を利用する方法論が主流となるはずである。そこで、この方法論に馴染むよう、あらかじめ VAE を設計しておく必要がある。そのため、2019 年度から、単語トークンごとの分散表現を利用できる VAE の提案を研究目的とした。

$$p(\mathbf{w}|\alpha,\beta) = \int_\theta \left(\prod_{n=1}^{N} p(w_n|\beta,\theta) \right) p(\theta|\alpha) d\theta.$$

図 2．既存のニューラルトピックモデル

３．研究の方法

(1) 2018 年度の研究方法

上述のように、2018 年度はまだ当初の研究目的に沿った研究を計画していた。そこで、トピックモデルにおける RNN の利用を目指すにあたって、二つの方法を採ることにした。

① RNN を利用するためには、まず、PyTorch などの深層学習フレームワークを使ってトピックモデルの変分推論を変分オートエンコーダ(VAE)として実装し、ハイパーパラメータのチューニングも含めて VAE 方式の推論計算に習熟する必要があった。この習熟のために、以下の方法を採用した。つまり、当時まだトピックモデルに使われていなかった敵対的変分ベイズ(AVB)[Mescheder+ arXiv:1701.04722]をトピックモデリングに使うと、元の VAE のように正規分布を使う場合と比べて変分推論の性能を改良できるかの検証を例題として、VAE 方式の推論計算に習熟する、という方法を採用した。

② さらに、深層学習フレームワークを使うことによる RNN の実装そのものに慣れる必要もあった。そこで、短歌自動生成を例題として設定し、約 14 万件の短歌を RNN に学習させ、短歌を自動生成する手法を提案、実装するという方法を採用した。

(2) 2019 年度の研究方法

２の(2)で述べたように 2019 年度から研究目的を変更した。つまり、文書中の各単語トークンについて分散表現を与えるニューラルネットワークを使うように作り直された VAE の提案を、研究の目的とした。トピックモデルのためのこのような VAE は、二つの研究方法の組み合わせで実現された。

$$\log p(\boldsymbol{w}|\alpha,\eta) \geq \sum_d \mathbb{E}_q \log \left[p(\boldsymbol{z}_d|\alpha) \prod_i \beta_{z_{di} w_{di}} \right] + \sum_k \mathbb{E}_q \log p(\boldsymbol{\beta}_k|\eta) + \mathcal{H}(q),$$

図 3．Mimno らの手法での ELBO

① 変分推論の実施にあたっては、各文書のトピック確率のほうを積分消去し、各単語トークンのトピック割り当てを表す離散潜在変数の方を残す、という方法を採った。この周辺化は、LDA の collapsed Gibbs sampling[Griffiths+ pnas.0307752101]で使われていた周辺化である。Mimno らは、この周辺化を変分ベイズ法でも用いた[Mimno+ arXiv:1206.6425]（図 3）。この Mimno らの研究アイディアを起点として、トピックモデル用の新しい変分推論を提案することを試みた。

② しかし、トピック割り当てを表す離散潜在変数の方を残した上で、変分推論を素直に実装してしまうと、トピック割り当ての事後確率を文書全体でまとめて推定する必要があり、組合せ論的爆発が生じる。Mimno らはサンプリングによる近似を提案しているが、本研究では、VAE と同じく amortized inference により、変分事後分布のパラメータをニューラルネットワークの出力として得るという方法を採ることにした。具体的には、Gumbel softmax trick で用いるパラメータを単語トークンごとに多層パーセプトロンの出力として得ることにした。

(3) 2020 年度の研究方法

最終年度は、完全に予想外の出来事があった。コロナ禍である。

2019 年度には、予定通り、単語トークンごとの分散表現を利用する VAE を提案することはできた。だが、この提案には２つ問題があった。第一に、事後分布を事前分布に近づける正則化項としての KL 情報量の項の近似に技術的な問題があった。実装自体はうまく動き、データセットによっては良い perplexity を得られていたものの、KL 情報量の項の計算は近似計算になっており、厳密に言えばこの近似を入れた ELBO が対数周辺尤度の下限である保証がないような近似方法になってしまっていた。第二に、提案手法は、離散潜在変数を周辺化したニューラルトピックモデルよりも良い perplexity を、実験で用いた三つのデータセットのうち一つでしか与えなかった。つまり、perplexity で評価した実際上の性能も思ったほど良くなかった。その理由は、ハイパーパラメータのチューニングが難しかったことにもよる。

2019 年度提案の手法がこのような問題を抱えていたため、思い切って、文書中の各単語トークンについて分散表現を与えるニューラルネットワークを使う VAE の案を、最初から考え直すことにした。しかし、この考え直しには、動機があった。

それは、Yoon Kim らが 2019 年に確率的文脈自由文法に関して提案したアイディア[Kim+ arXiv:1906.10225]を、トピックモデルの変分推論にも使ってみたいという動機である。

そこで、2020 年度は、モデルとしては LDA の原論文のものそのままを使い、原論文に書かれた ELBO に含まれる変分事後分布のパラメータを、原論文とは異なり、文書中の各単語トークンについて分散表現を与えるニューラルネットワークから得る、という研究方法を採ることにした。文書ごとのディリクレ事後分布のパラメータについては、原論文の変分ベイズ法と同様、単語トークンごとのトピック確率を同一文書内で集約することで得ることにした。

これによって、原論文のように微分してイコールゼロとおくような EM アルゴリズム式の ELBO 最大化ではなく、別の ELBO 最大化を実現できる。つまり、隠れ変数が個々の値を取る確率の更新と、事後分布のパラメータの更新を交互に繰り返すのではなく、隠れ変数が個々の値を取る確率をニューラルネットワークの出力として得て、これを使って他の推定すべきパラメータを表現し、そしてこれらを ELBO の式に当てはめることでニューラルネットワークの重みを勾配法で更新するのである。これは、文書中の単語トークンごとに与えられた分散表現の、新しい使い方の提案でもある。

この研究方法を実装し、評価実験を繰り返すことで、最終的には、文書中の各単語トークンについて前後の文脈に依存した分散表現を与えるニューラルネットワークとして RNN を実際に提案手法の中で使い、評価するつもりだった。しかし、コロナ禍の影響で計画は遅れた。文書内の各単語について分散表現を得るネットワークとしては、まだ多層パーセプトロン(MLP)を利用している段階で、年度末を迎えることとなった。

４．研究成果

(1) 2018 年度の研究成果

① 文書モデルでの AVB(adversarial variational Bayes)の利用（ICDPA2018 フルペーパー）：本研究課題は、トピックモデルにおける RNN の利用をテーマとしていた。その際、事後分布推定方法として変分ベイズ推定(VB)を使う。VB における近似事後分布の設定手法として、深層学習分野では主に変分オートエンコーダ(VAE)が使われる。VAE では対角正規分布が近似事後分布としてよく用いられ、そのパラメータを ELBO の最大化により求める。一方、より柔軟な近似事後分布を設定する手法として AVB（敵対的変分ベイズ)が 2017 年に Mescheder らによって提案された。これを文書モデリングに使い、柔軟な事後分布近似を実現した。

② RNN による短歌自動生成（ICCS2018 ポスター）：約 14 万件の短歌を RNN に学習させ、短歌を自動生成する手法を提案した。生成された短歌のスコア付けにはトピックモデルを使い、高スコアのものだけ出力する。この研究を通して、RNN の訓練に関する経験を蓄積できた。
③ 時間情報を利用した LDA のミニバッチ変分ベイズ（PRICAI2018 ショートペーパー）： LDA の VB に深層学習フレームワークを使うことはそれほどまだ広く行われていない。この研究では、トピック毎の単語確率分布に時間情報を反映させた LDA の推定計算を、PyTorch のテンソルのブロードキャストを利用して実現した。
④ トピックモデルでの AVB の利用（ADMA2018 ショートペーパー）：この研究は、①の継続で、AVB をトピックモデルの変分推定に利用した。これにより、トピックモデルにおいても AVB を柔軟な事後分布近似のために使えることが分かった。その結果、RNN を使ったトピックモデルへの AVB の適用可能性の感触を得た。

(2) 2019 年度の研究成果
① LDA（潜在的ディリクレ配分法）の変分ベイズ推定において、各書ごとのトピック確率を周辺化して amortized inference を行う手法の提案（国際会議 ICWE2019 併設ワークショップ）：前年度の「今後の研究の推進方策」において、次のような展望を述べた。つまり、各単語トークンのトピック割り当てを表す離散潜在変数を、LDA 向け VAE の従来研究とは異なり、周辺化して消去せずに利用するという展望である。今年度はこの展望を実行に移した。提案手法は 2 つのアイディアから成る。第一に、LDA の ELBO において各文書のトピック確率を積分消去し、各単語トークンのトピック割り当てを表す離散潜在変数の方を残す。これは Mimno らが 2012 年に提案したアイディアである。しかしトピック割り当ての事後確率を文書全体でまとめて素直に推定しようとすると組合せ論的爆発が生じる。そこで第二に、Mimno らのようにサンプリングでこの問題に対応するのではなく、amortized inference により、各トークンにおいて Gumbel softmax trick で用いるパラメータを同一のニューラルネットワークの出力として得た。こうして異なる単語間に関連性を持たせて、単語トークンごとのトピック確率をパラメータ化すると、データセットによっては前年度の実験結果よりも良い test perplexity が得られた。
② 上記①の手法を注意機構により拡張した手法の提案：上記の提案手法においてでは、各単語トークンについてのトピック割り当て確率を得る計算に、注意機構を導入することもできる。これにより、同一文書内のトークン間のトピック割り当て確率に依存関係をもたせるようなモデル化が可能となる。このアイディアの有効性を検証している途中で年度末を迎えたため、最終年度中にさらに実験を継続して進めることとした。

(3) 2020 年度の研究成果
本年度は、コロナ禍で計画通り研究が進まず、内容上も予想外の困難があった。だが、新しいアイディアにも辿り着いた。
① 予定では、LDA における Gumbel softmax trick の利用を継続するつもりだったが、トピック割り当てを表す離散潜在変数だけを残すような周辺化に基づくモデルでは、複数のデータセットで評価実験を繰り返すと、データセットによってハイパーパラメータのチューニングが非常に困難となることが分かり、なかなか良い perplexity を得られなかった。そこで、このアプローチは放棄し、一から推論手法を考え直すことにした。
② とはいえ、研究としての新規性を出すには、従来の neural topic model のように、文書のトピック確率を VAE のエンコーダの出力から得る手法をそのまま採用できない。そこで、同時進行で音声合成のための深層学習によるシーケンスデータのモデル化を試みる（ICTC2020 採録論文）など知見を広げつつ、試行錯誤した結果、次のアイディアに辿り着いた。つまり、各文書に含まれる単語の埋め込みのシーケンスを入力とするニューラルネットワークの出力として得た変分事後分布のパラメータを、VAE の枠組み抜きで ELBO 最大化にそのまま使う、というアイディアである。これは Yoon Kim らが 2019 年に確率的文脈自由文法に関して提案したアイディアを参考にしており、LDA の ELBO 最大化でも同様の amortization を行うことで、ある程度は良い perplexity が得られると分かった。検証実験はほぼ終わっており、近々論文化の予定である。ただし、元々の変分ベイズ推論よりも良い perplexity を得るには至っていないため、さらなる改良の余地は残っている。
③ この②のアイディアを実装するにあたっては、前年度利用を躊躇した RNN や注意機構についても、コードを変更するだけで対応できるようにしてある。つまり、期間全体の目標であった LDA における RNN の利用、しかも今までにない形での利用について、コードを少し変更するだけのところまで到達した。コードを動かし、ハイパーパラメータのチューニングをして性能を出すことは、期間内に間に合わなかったが、今後このまま実験を進める予定である。
さらに言えば、このように単語トークンごとの分散表現の列を出力するニューラルネットワークであれば何でも部品として使えるように、すでに ELBO 最大化のアルゴリズムを作成してあるため、BERT のような Transformer 系の言語モデルも利用可能である。この方向性は、後続の研究で追求する予定である。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Yuzana Win, Tomonari Masada	４．巻 1
２．論文標題 Myanmar Text-to-Speech System based on Tacotron-2	５．発行年 2020年
３．雑誌名 Proceedings of 2020 International Conference on Information and Communication Technology Convergence (ICTC)	６．最初と最後の頁 578-583
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICTC49870.2020.9289599	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Tomonari Masada	４．巻 11609
２．論文標題 Context-dependent Token-wise Variational Autoencoder for Topic Modeling	５．発行年 2020年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 35-47
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-51253-8_6	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomonari Masada, Atsuhiro Takasu	４．巻 10862
２．論文標題 LDA-Based Scoring of Sequences Generated by RNN for Automatic Tanka Composition	５．発行年 2018年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 395-402
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-93713-7_33	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomonari Masada, Atsuhiro Takasu	４．巻 11013
２．論文標題 Mini-Batch Variational Inference for Time-Aware Topic Modeling	５．発行年 2018年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 156-164
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-319-97310-4_18	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomonari Masada, Atsuhiro Takasu	４．巻 11323
２．論文標題 Adversarial Learning for Topic Models	５．発行年 2018年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 292-302
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-05090-0_25	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Tomonari Masada
２．発表標題 Context-dependent Token-wise Variational Autoencoder for Topic Modeling
３．学会等名 5th International Workshop on Knowledge Discovery on the Web (KDWEB 2019)（国際学会）
４．発表年 2019年

１．発表者名 Tomonari Masada
２．発表標題 Document Modeling with Implicit Approximate Posterior Distributions
３．学会等名 ICDPA 2018（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

ICCS 2018に論文がショートペーパーとして受理されました
http://diversity-mining.jp/wp/?p=535
PRICAI 2018に論文がショートペーパーとして受理されました
http://diversity-mining.jp/wp/?p=581
ADMA 2018に論文がショートペーパーとして受理されました
http://diversity-mining.jp/wp/?p=600

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
ミャンマー	Yangon Technological University			