ニューラルトピック将来予測モデルの構築

中村　礼音†　　森嶋　厚行††　　伊藤　寛祥††

† 筑波大学情報学群知識情報・図書館学類　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †nakamura.reon@klis.tsukuba.ac.jp, ††{mori,ito}@slis.tsukuba.ac.jp

あらまし　文書集合から潜在的なトピックを推定するモデルとしてトピックモデルが存在し，テキスト分析の分野において多くの成果が得られている．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．本研究では，時間情報が含まれる文書データを対象に，ニューラルネットワークを用いて，未観測の将来のトピックおよび各トピックの将来の文書量を予測するニューラルトピック将来予測モデルの構築を提案する．また，さまざまな手法で将来予測モデルを構築し，建築分野の特許文書と新型コロナウイルスに関するニュースデータを対象に手法の実験による検証結果を報告する．実験では，トピックモデルの推定とトピックの予測モデルの学習を同時に行うモデルが精度を維持したまま学習時間を削減可能であることがわかった．

キーワード　情報抽出，意味解析，時系列コーパス

1　序　　論

文書集合から潜在的なトピックを推定するモデルとしてトピックモデルが存在し，テキスト分析の分野において多くの成果が得られている．トピックモデルは，データに含まれる共起情報を扱った理解しやすいモデルであるため，「文書分類」「テキスト要約」「検索推薦」など，幅広い領域で利用されている [1] [2] [3]．潜在ディリクレ配分法（LDA）[4] は，単語の共起性を統計モデルとして扱った代表的なトピックモデルの手法である．具体的には，各文書には潜在的なトピックがあると仮定し，統計的に共起しやすい単語の集合が生成される要因を，この潜在トピックという観測できない確率変数で定式化している．LDA では，一つの文書には複数の潜在トピックが存在すると仮定し，そのトピックの配分率をベイズ推論（変分ベイズ推定・ギブスサンプリング）を用いてモデル化する．またニューラルトピックモデル（NTM）[5] [6] [7] は，変分オートエンコーダ（VAE）[8] を用いて潜在トピックとトピックの配分率を推論する手法である．NTM は多層パーセプトロンをもつニューラルネットワークモデルで学習を行っているため，LDA と比べてはるかに複雑な推論を行うことができる．

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」と「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．

ニューラルトピック将来予測モデルの基本的なアイデアは，「将来の潜在トピック」と「将来のトピックを構成する文書量」の予測に基づいているため，様々な学習方法や予測モデルを利用する方法が考えられられる．その中の一つの手法は，次のようなものである．はじめに，文書集合から話題となっているトピックとトピックを構成する文書量を把握するために，ニューラルトピックモデルを導入する．次に，「未観測の将来のトピック」と「将来のトピックを構成する文書量」を把握するための予測モデルとして LSTM [9] を導入する．予測のフェイズでは，NTM から推定されたトピックとトピックを構成する文書量を用いて，次のタームのトピックとトピックを構成する文書量を推定するための学習を行う．この 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．例えば，ここでニュース記事のような時系列文書集合を特定の期間ごとに分析した場合を仮定する．特定の期間ごとに意味解析を行った場合，トピックは入力コーパスに依存して生成されるため，それぞれの期間で全く意味の異なるトピックが生成される．そのため，生成された各期間のトピックと各期間のトピックの配分率を比較しても，トピックの変化やトピックのトレンドの推移を理解することができない．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．トピックモデルとして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 [4] 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：Mean Squared Error）を用いて予測の性能評価を行った．

本研究の貢献は次の通りである．

（1） 我々の知る限り，本論文はニューラルネットワークを用いて時系列文書集合を対象にトピックモデルの予測を行う「ニューラルトピック将来予測モデル」の構築を提案する初めての論文である．

（2） トピック将来予測モデルの構成要素として，時系列予測を可能にするための事前学習，トピック学習モデル，将来予測モデルの組み合わせという枠組みを提示する．

（3） 実世界データを対象に，いくつかの具体的なモデルを構築して実験評価している．その結果，トピックモデルの推定とトピックの予測モデルの学習を同時に行う手法が，予測精度を維持したまま学習時間を削減できることを明らかにした．

2 関 連 研 究

トピックモデルは，文書が生成される過程を，確率を用いてモデル化した確率生成モデルであり，確率論の枠組みでさまざまな情報を統合できるため，多様な情報を扱うためのトピックモデルの拡張が数多く提案されてきた．しかし，図 1（本研究と特に関係が深い既存手法と提案手法を比較した表）に示すように既存手法には，時系列コーパスを用いた予測モデルはほとんど存在しない．1 節，2 節で本研究に関連したさまざまなトピックモデル手法を概説し，提案手法との関係や違いを述べる．次に 3 節で，トピックの変化をとらえるための予測モデルを導入する．

2.1 静的コーパスのためのトピックモデル手法

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．トピックモデルの最も代表的な手法である潜在ディリクレ配分法（LDA）[4] は，トピックを生成する分布とトピックの単語を生成する分布にディリクレ事前分布を仮定し，ベイズ推定（変分ベイズ推定やギブスサンプリング）する手法である．LDA を拡張した手法もさまざま提案され，その有用性が確認されてきた．しかしながら LDA は，新しいデータを用いて推論をする際に，モデルをあらためて構築する必要がある．また，データ数が増しより複雑な推論が必要とされるにつれて，これらのパラメータの高速かつ正確な推論を行うことが難しくなってきている．

	精度	時系列コーパス	予測
LDA	△	×	×
DTM	△	○	×
TM-LDA	△	△	△
NTM	○	×	×
提案手法	○	○	○

図 1: 既存手法と提案手法の比較.

Wang らによって提案されたニューラルトピックモデル（NTM）[5] [6] [7] は，これらの問題を解決した．NTM は，変分オートエンコーダ（VAE）[8] を利用したトピックモデルであり，入力データをおおよそ多変量標準正規分布に従うランダムな潜在変数に次元削減させる．さらにランダムな潜在変数からサンプリングし，多層パーセプトロン層（MLP）をもつニューラルネットワークモデルを経由することで入力を再構成する生成モデルである．NTM は MLP をもつニューラルネットワークモデルであり，LDA と比べてはるかに複雑な推論が可能となった．また NTM では，並列計算や GPU を用いた実装も可能となり，高速な推論も行うことができる．損失関数を同時に最適化するだけでほかのモデルとの融合学習も容易に行うことができるため，本研究では NTM を利用することとする．

2.2 時系列データを扱ったトピックモデル手法

時間情報を考慮したトピックモデルは，これまで数々提案されてきた．もっとも代表的な例として，M.Blei らによる Dynamic Topic Model（DTM）[10] がある．DTM では，トピックごとの単語部出現確率が変化することを仮定する．単語分布は連続値をとる確率変数であるため，連続型の確率分布である正規分布を生成分布とすることで，以前の状態のパラメータを受け継いだ意味解析を行うことができる．DTM を用いることで，トピックの様々な潜在情報の変化を扱うことが可能となった．しかし，DTM は時間情報を考慮することにのみ着目しており，将来のトピックとトピックのトレンドの予測を行うことはできない．そのほかに時間情報を扱ったトピックモデルとして，そのほかにも J.Rieger らの Rolling LDA [11] や H.Amoualian らの Streaming-LDA [12] などが挙げられるが，こちらも DTM と同様に将来のトピックとトピックのトレンドの予測を行うことはできない．

将来のトピックの分布を予測する数少ない例として，Y.Wang らが提案した Temporal-LDA（TM-LDA）[13] がある．TM-LDA は，同一著者による一連の投稿のようなテキストストリームを効率的にマイニングするためのトピックモデル手法であり，学習された TM-LDA を用いることで将来の投稿におけるトピック分布を予測することができる．TM-LDA は，7 日前までの文書コーパスを持ちることでその期間のトピックの分布を生成し，8 日目のトピックの分布を予測するモデルである．そのため，TM-LDA は各期間ごとで得られるトピックに整合性がなく時間遷移を考慮した分析を行うことはできない．また，トピックの単語分布の予測は行うことができない．

さらに，これらの時系列データを扱ったトピックモデル手法

はいずれも LDA に基づいて構築されたものであり，NTM を使用した時間情報と予測を扱ったトピックモデル手法は，執筆者の調べたところ存在しない．

2.3 予測モデル

トピックモデルは，文書の大まかな意味構造をとらえることはできるが，時間ごとのトピックの変化をとらえることは困難である．そのため，本節ではトピックの変化をとらえるための予測モデルをいくつか紹介し，本研究で用いる予測モデルを導入する．

時間におけるデータの変化をとらえるための統計モデルを，時系列解析と呼ぶ．時系列解析は回帰分析の一種で，目的変数に現在の値を，説明変数に過去の値を設定することで関係を数式化する．代表的な時系列解析手法として，自己回帰（AR）モデルと移動平均（MA）モデルが用いられてきた．AR・MA モデルから派生したモデルとして，自己回帰平均移動（ARMA）モデル，自己回帰和分平均移動（ARIMA）モデルなどが提案された．

一方で，モデルに高い表現力が求められるにつれ，ARMA モデルなどの線形モデルでは十分な予測精度を得ることは困難になってきた．再帰型ニューラルネットワーク（RNN）は，再帰構造を持つ多層のニューラルネットワークをシーケンスごとにつなぎ合わせることで，時系列データの複雑な学習を可能としている．RNN の中でも，Long Short Term Memory（LSTM）[9] は，多種多様な問題で高い精度を出しており，広く用いられている．本研究では，高い予測精度を持ちかつ，ニューラルトピックモデルとの融合学習も行いやすい LSTM を予測モデルとして用いる．

3 前提知識

本章では，本研究で用いるニューラルトピックモデルのアーキテクチャと数学的な定義を説明する．

3.1 Neural Topic Model

本研究では，変分自己符号化器（VAE）を用いて潜在トピックを抽出するニューラルトピックモデル（NTM）を使用している．VAE とは，次のようなプロセスで x を再構成する仕組みである．はじめに入力ベクトルを与えると条件付き分布 $p(z|x)$ に従って変数 z を生成する．次に生成された z を与えると条件付き分布 $p(x'|z)$ に従って再構成ベクトル x を生成する．このように z を生成する前者のプロセスを Encoder，x を生成する後者のプロセスを Decoder と呼ぶ．トピックモデルに置き換えた場合，トークンの Bag of Wards（BoW）ベクトルである x_{bow} が入力ベクトルとなり，Encoder と Decoder のプロセスを経て，再構成された BoW ベクトル x'_{bow} を求めることと同一である．ここで x は観測することができる変数である一方，z は観測することができない変数である．本論文では z を他の変数と区別するために潜在変数 z と呼ぶこととする．

はじめに Encoder のプロセスを説明する．上の説明にて，変数 z が分布 $p(z|x)$ から生成されることを確認した．ここで $p(z|x)$ がおおよそ標準正規分布 $q(z)$ に従うと仮定する．このとき $p(z|x)$ のパラメータ　と log　は，x_{bow} を入力としたニューラルネットワークモデルで学習される．以下が　と log　の導出過程である．

$$= f\ (f_e(x_{bow})); \quad \log \quad = f\ (f_e(x_{bow})) \tag{1}$$

ここで $f\ (\ )$ は ReLU 活性化関数によるニューラル層である．

潜在変数 z は以下のようにサンプリングされる．

$$z \quad N(\ ;\ ^2) \tag{2}$$

続いて Decoder のプロセスを説明する．Decoder のプロセスでは，変数 z から再構成ベクトル x' を生成する．具体的には，サンプリングされた変数 z を入力としたニューラルネットワークモデルが，再構成ベクトル x' を出力する．数式で表現すると，以下のように変数 z から x' が導出される．

$$= softmax(f\ (z)) \tag{3}$$

$$x'_{bow} \quad softmax(f\ (\ )) \tag{4}$$

Encoder と同様に，$f\ (\ )$ は ReLU 活性化関数によるニューラル層である．

既存のトピックモデル同様出現語彙の個数を V，トピックの個数を K としたとき，式 (3) の　が K 個のトピック分布，式 (4) のパラメータ　がトピックごとの V 個の単語分布となる．

4 提案手法

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」と「将来のトピックを構成する文書量」を予測するための「ニューラルトピック将来予測モデル」の構築を目指す．最適な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．本章では，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する融合学習モデルの構築方法を紹介する．また本章で紹介するモデルの変種として，同時に最適化を行わず順に最適化を行うモデルが考えられるが，同様の議論を行うことで導出することができる．

4.1 予測モデル

はじめに予測モデルで使用する，Θ を導出する．前節では，文書ごとのトピック分布を　と表現した．　は期間 t 中に出現した文書数だけ存在する．しかし予測モデルでは各期間全体でのトピックのボリューム変化を捉えることが目的であるため，各期間ごとのトピックの配分率を以下のように表現する．

$$\Theta = \sum_{n=1} {}_k \tag{5}$$

ここで Θ は期間 t 中の文書ごとのトピック分布の総和とした．なお現在のトピックの配分率 Θ_t から 1 ターム先のトピックの配分率 Θ_{t+1} を予測する際に，入力するデータをスケールする必要がある．そのため各期間ごとに平均して含まれるコーパス

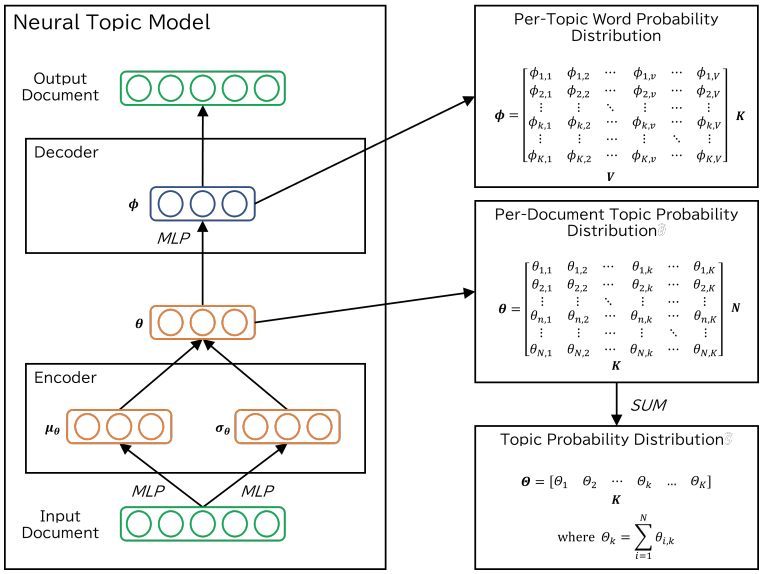

図 2: NTM のアーキテクチャ．

数で割ることによって，データのスケールを確保した．

$$\Theta_{scaled} = \frac{\Theta}{\bar{N}} \tag{6}$$

ただし $\bar{N}$ は，各期間ごとに含まれる平均コーパス数である．

続いて予測モデルの作成プロセスを紹介する．単純なフィードフォワード・ニューラルネットワークでは，時系列データの性質を十分に学習することができないため再帰型ニューラルネットワーク（RNN），中でも LSTM を使用する．予測のフェイズでは現在の Θ_t と $\quad_t$ を LSTM の入力とし，1 ターム先の Θ_{t+1} と $\quad_{t+1}$ を予測する．予測は，各期間ごとで Θ_{t+1} と $\quad_{t+1}$ を入力し予測モデルのパラメータを更新するオンライン学習の方法を利用する．トピックの配分率の予測モデルとトピックの単語分布の予測モデルは定式化される．

$$\hat{\Theta}_{t+1} = f\ (\Theta_t; h_t) \tag{7}$$

$$\hat{\quad}_{t+1} = f\ (\quad_t; h_t) \tag{8}$$

4.2 モデルのアーキテクチャ

本節では，ニューラルトピックモデル（NTM）と LSTM を融合学習するプロセスを紹介する．はじめに，t 期のコーパスの Bag of Words 表現されたベクトルを NTM の入力として与えることで，t 期のトピックの配分率 Θ_t とトピックごとの単語分布 $\quad_t$ を推定する．次にこれらの生成分布 Θ_t と $\quad_t$ をそれぞれ予測モデル（LSTM）に与えることで，1 ターム先のトピックの配分率 Θ_{t+1} とトピックごとの単語分布 $\quad_{t+1}$ を予測する．ここで使用された NTM と Θ と $\quad$ の LSTM の損失関数を同時に誤差逆伝播法によって最小化する．以上の流れを各期間ごとに行うことモデルを訓練する．モデルのアーキテクチャは図 3 に示してある．

なお，NTM は時間情報を持たない静的なコーパスを分析することを前提としており，異なる期間のコーパスで意味解析を行った際，潜在的なトピックの内容が全く異なってしまう可能性がある．しかしトピックのトレンドの遷移を正しく理解するためには，異なる期間であっても潜在的なトピックの内容は一貫している必要がある．この問題を解決するために事前学習済み NTM を利用する．事前学習済み NTM は，あらかじめ訓練データとなる期間の文書集合全体でニューラルトピックモデルを訓練することによって構築する．次にそれぞれの期間でニューラルトピックモデルを訓練する際，事前学習済み NTM のパラメータ全体を各期間のモデルに受け継ぐことによって，異なる期間でも整合性のあるトピックを生成できるようにする．

4.3 モデルの訓練

各モデルの損失関数を定義する．ニューラルトピックモデル（NTM）の場合，目的関数は負の変分下限に基づいて以下のように定義される．

$$L_{ntm} = D_{KL}(p(z)jjp(zjx)) \quad E_{q(xjz)}[p(xjz)] \tag{9}$$

損失関数の第一項は z を生成する分布 $p(zjx)$ と多変量標準正規分布 $p(z)$ との KL ダイバージェンス損失であり $p(z)$ に近似するよう $p(zjx)$ が学習される．続いて損失関数の第二項は交差エントロピー再構成損失であり，入力 x_{bow} と出力 x^0_{bow} が一致するように学習される．NTM の損失関数は，第一項が Encoder の損失関数，第二項が Decoder の損失関数を表す．次に，予測モデル（LSTM）の目的関数には平均二乗誤差（MSE）を導入する．

$$L_{lstm} = \frac{1}{n}\sum_{i=0}^{n}(\hat{\Theta}_i \quad \Theta_i)^2 \tag{10}$$

$$L_{lstm} = \frac{1}{n}\sum_{i=0}^{n}(\hat{\quad}_i \quad _i)^2 \tag{11}$$

次に NTM と LSTM の損失関数を線形結合して誤差逆伝播法によって最小化することで，NTM と LSTM を同時に最適化することができる．損失関数は

$$L_1 = L_{ntm} + L_{lstm} \tag{12}$$

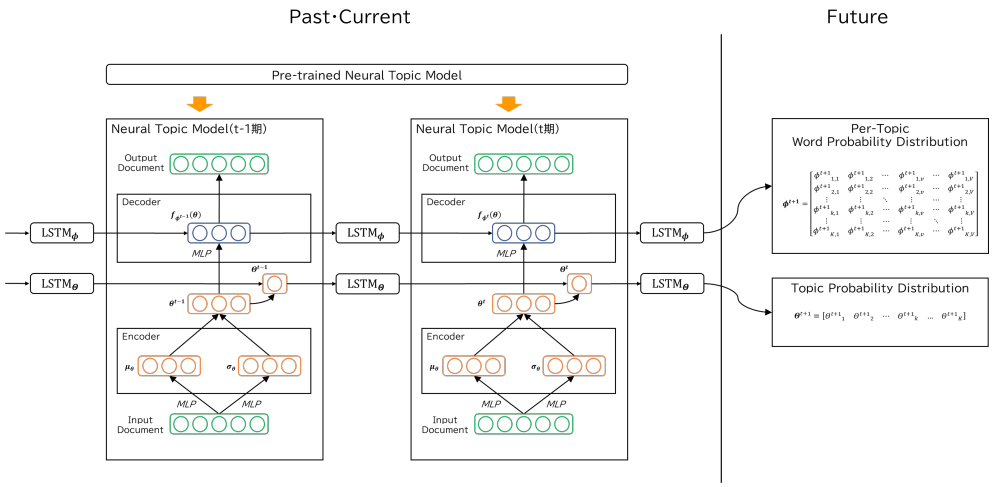

図 3: 提案手法のアーキテクチャ．

$$L_2 = L_{ntm} + L_{lstm} \tag{13}$$

のように Θ と　の場合それぞれで作成し，特定のエポックごとに Θ と　の最適化を行う．以上のように 2 つのモデルのパラメータを同時に最適化することによって，NTM と LSTM の融合学習を行う．また，実験では 2 つのモデルの損失関数を別々に最適化した 2 フェイズのモデルも同様に性能評価を行う．

5　実　　験

実験では，「建築分野の特許文書」と「新型コロナウイルスに関するニュース記事」をデータセットとして使用した．本研究で提案している手法は，トピックモデルと予測モデル双方を利用したモデルであり，それぞれの性能評価が必要となる．トピックモデルとしての性能評価指標として Perplexity，予測モデルの性能評価指標として二乗誤差（MSE）を導入し，既存のモデルと比較することで，学習の有効性を検証した．

5.1　データセット

本研究では，日本語の時系列コーパスとして (1)「建築分野に関連した特許文書（1999 年から 2019 年）」と (2)「Yahoo ニュースに掲載された新型コロナウイルスに関するニュース記事（2020 年 6 月から 2022 年 2 月）」を使用した．両コーパスとも形態素解析ソフトウェアの Mecab と mecab-ipadic-NEologd 辞書を用いてパース処理を実施した．はじめに，「建築分野の特許文書」に関するパース処理の過程を説明する．「建築分野の特許文書」には，「前記」「図」「先行文献」などといった文書に共通のフレーズが含まれている．そのため，頻出度が上位 100 位までの単語はあらかじめ取り除いた．また，全文書中に含まれる単語のうち出現回数が 5 回以下の単語は除去し，特殊記号や数値は取り除いた．次に，「新型コロナウイルスに関するニュース記事」は，全文書中に含まれる単語のうち出現回数が 3 回以下の単語は除去し，特殊記号や数値は取り除いた．これらの処理を図って，最終的に得られた各総単語数は (1)「建築分野の特許文書」: 69618 語，(2)「新型コロナウイルスに関するニュース記事」: 67801 語となった．

次にこれらの時系列コーパスを特定の期間ごとに分割した．(1)「建築分野に関連した特許文書（1999 年から 2022 年）」は，年ごとにデータセットを分割し 1999 年から 2017 年までを訓練データ，2018 年と 2019 年までをテストデータとして使用した．(2)「Yahoo ニュースに掲載された新型コロナウイルスのニュース記事（2020 年 6 月から 2022 年 02 月）」は，月ごとにデータセットを分割し 2020 年 6 月から 2022 年 7 月までを訓練データ，2022 年 1 月から 2022 年 2 月までをテストデータとして使用した．

5.2　性能評価指標

本研究で提案している手法は，トピックモデルと予測モデルを融合したモデルであり，それぞれの性能評価が必要となる．トピックモデルとしての性能評価指標として Perplexity，予測モデルの性能評価指標として二乗誤差（MSE）を導入し，既存のモデルと比較することで，学習の有効性を検証した．

はじめにトピックモデルとしての性能評価指標として用いる，Perplexity を説明する．Perplexity とは，言語モデルの汎化性能を図る性能評価指標であり，

$$Perplexity = \exp^{-\frac{\sum_{d=1}^{M} \mathcal{L}[w \mid M]}{\sum_{d=1}^{M} n_d}} \tag{14}$$

のように定義される．Perplexity は直観的には単語をどれだけ絞り込めたかを示した評価指標である．例えば，ある文書の 1 単語が隠されているとする．語彙数が 10000 のとき，ランダムなモデルでは，隠されている単語の選択肢の数は 10000 となる．例えば，あるモデルの Perplexity が 1000 の場合，それは隠された単語の選択肢の数を 1000 まで単語の候補を絞ることができたことを示す．つまり，Perplexity がより低いモデルほど，よりよいモデルであるといえる．

続いて，予測モデルの性能評価指標として，平均二乗誤差（MSE）を導入する．以下のような数式で定義される．

$$MSE = \frac{1}{n} \sum_{i=0}^{n-1} (\hat{y}_i - y_i)^2 \tag{15}$$

予測値と実測値が近ければ近いほど，MSE の値は低い値となる．

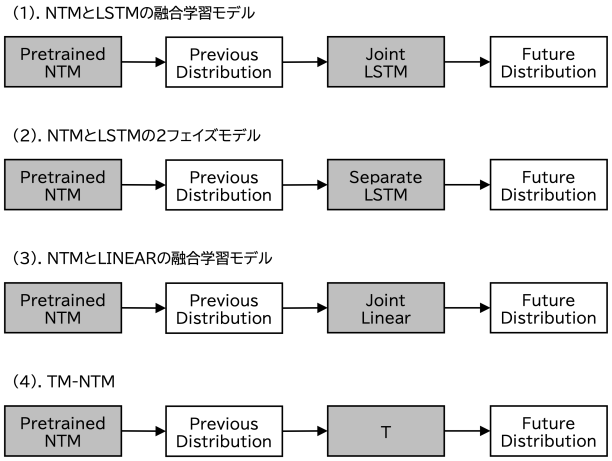

図 4: 比較手法.

5.3 実験設定

次に，実験の比較方法を説明する．はじめに，予測モデルとの性能評価と融合学習の効果を検証する．検証のために以下の 4 つのモデルを導入する．いずれのプロセスも Train データから構築された Pre-Training NTM を用いて，新しいデータのトピックとトピックを構成する文書量を推定するという点では一致している．

(1) NTM と LSTM の融合学習モデル；ニューラルニューラルトピックモデル（NTM）と LSTM の損失関数を同時に最適化したモデル.

(2) NTM と LSTM の 2 フェイズ学習モデル；NTM の損失関数をはじめに最適化したのち，LSTM の損失関数を最適化する方法.

(3) NTM と LINEAR の融合学習モデル；NTM と単純な線形（Linear）モデルの損失関数を同時に最適化したモデル.

(4) TM-NTM（NTM と遷移パラメータ T の 2 フェイズモデル）；NTM と TM-LDA [13] の遷移パラメータ T を用いて予測を行うモデル.

これらのモデルの「Θ の予測モデルの平均二乗誤差（MSE）損失」「 の予測モデルの平均二乗誤差（MSE）損失」「Perplexity」を比較することで，学習の有効性を検証した．また，平等な比較をするために，それぞれのモデル隠れ層のサイズとニューラルネットワーク層のサイズは統一して検証を行った．なお，(4) TM-NTM はトピックの単語分布の予測はできないため，「Θ の予測モデルの平均二乗誤差（MSE）損失」のみ出力する．TM-LDA とは異なり，トピック分布の推定を NTM を用いて行っている．学習回数であるエポック数は，「特許文書」を 200，「ニュース記事」を 150 とした.

次に，既存のトピックモデル手法と比べたときの学習の有効性を検証するために，(5) 潜在ディリクレ配分法（LDA）を導入した．こちらも (1)(2)(3) と同様に，エポック数を「特許文書」: 200，「ニュース記事」: 150 で実験を行った．しかし，LDA では将来のトピックの予測を行うことはできないため，訓練期間のみの Perplexity を比較することで，トピックモデルとしての学習の有効性を検証した.

5.4 実験結果と考察

5.4.1 既存手法との Perplexity の比較

本節では，既存手法と本研究で用いているトピックモデルの学習法の精度を検証する．はじめに，(1)(2)(3)(5) の「Θ の「Perplexity」を比較する．「特許文書」をもちいたときと，「ニュース記事」をもちいたときの結果をそれぞれ表 1 と表 2 にまとめた．なお，ここでの「Perplexity」は学習期間それぞれでえられた Perplexity の平均である．(1)(2)(3) はニューラルネット層のサイズと学習回数のニューラルトピックモデル（NTM）を実行しているため，Perplexity の値はほとんど変わらなかった．(4)LDA の Perplexity の結果と (1)(2)(3) の NTM を用いたモデルをもちいた結果を比較してみると，NTM を用いたモデルの方がよりよい Perplexity を得ることが確認できた.

表 1: 「特許文書」を用いたときの Perplexity の比較

モデルの種類	Perplexity	計算時間
(1) NTM と LSTM の融合学習モデル	2466.5	102.8
(2) NTM と LSTM の 2 フェイズ学習モデル	2462.6	110.1
(3) NTM と LINEAR の融合学習モデル	2464.4	85.0
(5) 潜在ディリクレ配分法（LDA）	4735.6	148.5

表 2: 「ニュース記事」を用いたときの Perplexity の比較

モデルの種類	Perplexity	計算時間
(1) NTM と LSTM の融合学習モデル	912.5	141.8
(2) NTM と LSTM の 2 フェイズ学習モデル	912.8	149.1
(3) NTM と LINEAR の融合学習モデル	913.1	107.1
(5) 潜在ディリクレ配分法（LDA）	2669.9	211.2

5.4.2 融合学習の効果と予測精度

本節では，NTM と LSTM を融合学習することの効果と予測モデルとして LSTM を用いる意義を確認する．テストデータのトピックまたはトピックを構成する文書量を予測した際の平均二乗誤差（MSE）の値を，表 3 と表 4 にまとめる．ここでラベルとなっている Θ と　は NTM を訓練の時と同等のエポック数で学習して得られたものを用いた．仮説では，「(1) NTM と LSTM の融合学習モデル」の方が，「(2) NTM と LSTM の 2 フェイズ学習モデル」に比べてより低い MSE が得られると期待したが，両モデルの値を比較してみると有意な差はなかった．しかし，融合学習モデルは，2 フェイズモデルより短い実行時間でほぼ同等の精度を得ることができた．また，「(4) TM-NTM（NTM と遷移パラメータ T の 2 フェイズモデル）」に比べて，(1)(2)(3) のモデルはかなり良い値を得ることができ，ニューラル層を用いた予測の有効性が確認できた．さらに，「(3) NTM と LINEAR の融合学習モデル」と比べて，(1)(2) のモデルはより良い値を得ることができた．この結果から予測モデルとして，LSTM を用いたことの意義が確認できた.

最後に，実際のトピックの単語分布と予測モデルからえられたトピックの単語分布を出力した結果を紹介する．テストデー

表 3: 「特許文書」を用いたときの予測精度の比較

モデルの種類	MSE	MSE
(1) NTM と LSTM の融合学習モデル	0.000027	0.002208
(2) NTM と LSTM の 2 フェイズ学習モデル	0.000028	0.002219
(3) NTM と LINEAR の融合学習モデル	0.000069	0.10399
(4) TM-NTM	0.449250	-

表 4: 「ニュース記事」を用いたときの予測精度の比較

モデルの種類	MSE	MSE
(1) NTM と LSTM の融合学習モデル	0.00102	0.00311
(2) NTM と LSTM の 2 フェイズ学習モデル	0.00096	0.00316
(3) NTM と LINEAR の融合学習モデル	0.00276	0.22511
(4) TM-NTM	1.45295	-

タである 2016 年の「建築分野の特許文書」を使用し，実際のトピックの単語分布を図 5 に，予測モデルからえられたトピックの単語分布を図 6 にまとめた．なお，表示している単語数は上位 10 語としている．

5.4.3 結果の考察

続いて結果の考察を行う．本研究で提案した手法は，LDA と比べてよりよい性能を得ることができた．また，NTM と LSTM の損失関数を同時に最適化することによって，既存の NTM よりもモデルの性能が劣ることはなかった．実験を通じて，NTM を用いてトピックとトピックを構成する文書量の時間変化を観察することができた．

一方，予測性能は今回の結果のみを用いて精度を比較することは困難であったと考える．理由の一つが予測するための期間が少なかった点が挙げられる．具体的には，「建築分野の特許文書」は予測期間が 1999 年から 2015 年までの 17 ターム，「Yahoo ニュースに掲載された新型コロナウイルスのニュース記事」は 2020 年 6 月から 2021 年 12 月までの 19 タームしかなかった．十分なデータ数を確保して，100 タームほどの予測期間のあるデータセットでも実験を行うべきだと考えられる．

6 結　　論

本研究は NTM と予測モデルの融合学習によって，未観測の将来のトピックと将来のトピックを構成する文書量を予測することを目指した．

実験では，トピックモデルの性能と予測モデルの性能を，それぞれ Perplexity と平均二乗誤差（MSE）を性能評価指標として用いて検証した．提案手法は，既存手法と比べてよりよい性能を得ることができた．また，トピックモデルの推定とトピックの予測モデルの学習を同時に行う手法が，予測精度を維持したまま学習時間を削減できることを明らかにした．

一方，予測性能は今回の結果のみを用いて精度を比較することは困難であったと考える．理由の一つが予測するための期間が少なかった点が挙げられる．具体的には，「建築分野の特許文書」は予測期間が 1999 年から 2015 年までの 17 ターム，「Yahoo ニュースに掲載された新型コロナウイルスのニュース記事」は 2020 年 6 月から 2021 年 12 月までの 19 タームしかなかった．そのため，十分なデータ数を確保して，実験を行うべきだと考えられる．以上より，引き続き実験の検証が必要だと考えられる．

謝　　辞

本研究の一部は　JSPS 科研費 (22H00508，20K23337，22K17944)　および株式会社熊谷組の支援による．

文　　献

[1] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, pp. 29{41, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] Shixia Liu, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol.*, Vol. 3, No. 2, feb 2012.

[3] Miha Pavlinek and Vili Podgorelec. Text classi cation method based on self-training and lda topic models. *Expert Systems with Applications*, Vol. 80, pp. 83{93, 2017.

[4] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, Vol. 14. MIT Press, 2001.

[5] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference, 2017.

[6] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

[7] Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R. Lyu, and Shuming Shi. Topic-aware neural keyphrase generation for social media language, 2019.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[9] Sepp Hochreiter and J■rgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, p. 1735{1780, nov 1997.

[10] David M. Blei and John D. La erty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 113{120, New York, NY, USA, 2006. Association for Computing Machinery.

[11] Jonas Rieger, Carsten Jentsch, and J●rg Rahnenf■hrer. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2337{2347, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 695{704, New York, NY, USA, 2016. Association for Computing Machinery.

[13] Yu Wang, Eugene Agichtein, and Michele Benzi. Tm-lda: E cient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, p. 123{131, New York, NY, USA, 2012.