Sentence Embeddingsを導入した潜在トピックモデル

山本眞大[1] 萩原将文[1]

[1] 慶應義塾大学大学院 理工学研究科

1 はじめに

Latent Dirichlet Allocation (LDA) [1] は，各文書に複数の潜在トピックが存在すると仮定し，単語の集合が生成される要因を潜在トピックにより定式化する．潜在トピックモデルは，統計的機械学習の主要な研究分野の一つとなっており，様々な改良モデルが提案されている．その中でも近年の表現学習の進歩に伴い，潜在トピックモデルと単語の分散表現 (Word Embeddings) を組み合わせる研究が存在する [2, 3]．これらの研究では，単語の分散表現の集合が生成される要因を潜在トピックにより定式化する．これにより，分散表現が類似している単語同士が同じトピックに割り当てられやすくなり，意味的に一貫したトピックが生成されると同時にモデルの性能が向上することが示されている．

一方，表現学習の分野では単語の分散表現を獲得する研究だけではなく，文レベルの分散表現 (Sentence Embeddings) を獲得する研究も行われている [4, 5]．単語の分散表現がトピックモデルの性能向上に寄与することから，文の分散表現を同時に考慮することで更なるモデルの性能向上が見込めると考えられる．しかしながら，Sentence Embeddingsを陽に利用した潜在トピックモデルは提案されておらず，その有効性について明らかになっていない．

そこで本稿では，Word Embeddings と Sentence Embeddingsを同時適用する前段階として，Sentence Embeddingsを導入した潜在トピックモデルを提案し，その有効性について検証を行う．具体的には，各文のトピック分布を基に Sentence Embeddingsが生成されると仮定したモデルを提案する．これにより，文単位の意味表現を考慮したトピック割り当てが行われることが期待される．評価実験では，推定されたトピック分布を用いて文書分類精度の比較が行われた．その結果，Sentence Embeddingsを用いることで，文書分類の精度が向上することが確認された．また，文の分散表現化手法についての比較も行った結果，Skip-Thought Vectors により得られた分散表現を用いた場合の精度が一番高いことが確認された．

本稿の貢献は以下の通りである．

- 潜在トピックモデルにおいて Sentence Embeddings を導入したモデルを初めて提案したこと．
- 文書分類タスクの精度を比較した結果，Sentence Embeddingsを導入することで精度が向上することを示したこと．
- 文の分散表現化手法についての比較を行い，Skip-Thought Vectors により得られた分散表現を用いた場合に文書分類の精度が一番高くなることを示したこと．

2 関連研究

2.1 Latent Dirichlet allocation (LDA)

LDA は Blei ら [1] によって提案された，文書が潜在的なトピックから確率的に生成されると仮定するモデルである．LDA のグラフィカルモデルを図 1(a) に示す．LDA は文書中の単語を bag-of-words で表現し，各単語がどのトピックから生成されたのかを推定する．推定されたトピック分布は文書分類などのタスクの特徴量として使用可能である．

LDA は自然言語処理だけではなく多様な分野に適用可能なモデルであり，様々な改良モデルが提案されている．その中でも特に，LDA において単語の分散表現を用いる研究について 2.2 節で説明を行い，提案モデルの基となった Supervised topic models について 2.3 節で説明する．

2.2 LDA + Word Embeddings

近年の表現学習の進歩に伴い，潜在トピックモデルと単語の分散表現 (Word Embeddings) を組み合わせ

る研究が存在する [2,3]．これらの研究では単語の分散表現の集合が生成される過程をモデル化する．これにより，分散表現が類似している単語同士が同じトピックに割り当てられやすくなり，意味的に一貫したトピックが生成されやすくなることが報告されている．

潜在トピックモデルにおいて単語の分散表現の有効性が示唆される一方で，文レベルの分散表現は着目されておらずその有効性は明らかになっていない．本稿では，潜在トピックモデルにおいて Sentence Embeddings の適用を行い，その有効性の検証を行う．

2.3 Supervised topic models

Supervised topic models [6] は，LDA と正規回帰モデルを組み合わせたモデルであり，潜在トピックの学習に教師情報を利用することができる．具体的には，各文書の潜在トピックの集合を基に教師情報が生成されると仮定する．

提案モデルではこのモデルを参考に，各文のトピック分布を基に Sentence Embeddings が生成されるモデルを構築する．両者の違いは，Supervised topic models が文書単位のトピック分布を基に 1 次元の教師情報を生成するのに対し，提案モデルでは，文単位のトピック分布を基に多次元の Sentence Embeddings を生成することである．

3 Sentence Embeddings を導入した潜在トピックモデル

3.1 生成過程

提案モデルと従来の LDA の大きな違いは，トピック推定の際に文の分散表現を用いるか否かである．表 1 に提案モデルで用いられるパラメータを示す．これらのパラメータを用いて，提案モデルによる生成過程は以下の通りに表される．

1. For $k = 1$ to K
   - **(a)** Draw a topic word distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$
2. For each document d
   - **(a)** Draw a document topic distribution $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$
   - **(b)** For each sentence s in document d
     - **(I)** For the i-th word $w_{d,s,i}$ in sentence s
       - **(i)** Draw its topic assignment $z_{d,s,i} \sim \text{Categorical}(\boldsymbol{\theta}_d)$
       - **(ii)** Draw a word $w_{d,s,i} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{d,s,i}})$
     - **(II)** Draw a sentence vector $\boldsymbol{e}_{d,s} \sim \mathcal{N}(\boldsymbol{e}_{d,s}|\boldsymbol{\eta}^{\mathrm{T}}\bar{\boldsymbol{z}}_{d,s}, \sigma^2\boldsymbol{I})$

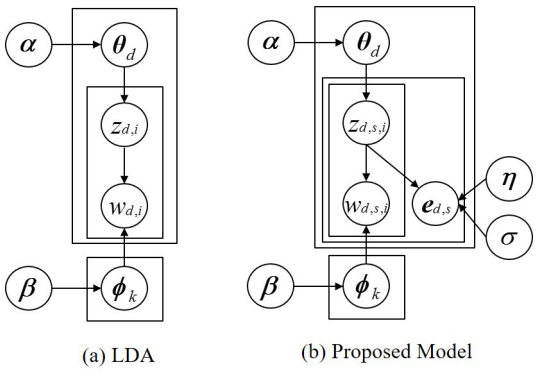

図 1: LDA と提案モデルのグラフィカルモデル

ここで，$\boldsymbol{I}$ は単位行列を表す．$\bar{\boldsymbol{z}}_{d,s}$ は文書 d 中の文 s のトピックを単語数で正規化したベクトルであり，以下の式で表される．

$$\bar{\boldsymbol{z}}_{d,s} = \frac{1}{n_{d,s}} \cdot \sum_{i=1}^{n_{d,s}} \boldsymbol{z}'_{d,s,i} \tag{1}$$

図 1(b) に提案モデルのグラフィカルモデルを示す．LDA と比較すると，提案モデルは潜在トピックを基に Sentence Embeddings を生成する部分が異なっていることが分かる．

3.2 事後分布の推定

提案モデルの事後分布を解析的に求めることは困難である．そこで，周辺化ギブスサンプリングを用い事後分布の近似分布を求める．サンプリングの際に必要となる条件付き確率は以下の通りである．

$$p(z_{d,s,i} = k|\boldsymbol{z}^{-d,s,i}, \boldsymbol{w}, \boldsymbol{e}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2)$$
$$\propto (n_{d,k}^{-d,s,i} + \alpha_k) \cdot \frac{n_{k,v}^{-d,s,i} + \beta_v}{n_{k,\cdot} + \sum_v \beta_v} \cdot \mathcal{N}(\boldsymbol{e}_{d,s}|\boldsymbol{\eta}^{\mathrm{T}}\bar{\boldsymbol{z}}_{d,s}, \sigma^2\boldsymbol{I}) \tag{2}$$

ここで，$n_{d,k}^{-d,s,i}$ および $n_{k,v}^{-d,s,i}$ は，それぞれ $\delta(z_{d,s,i} = k)$ および $\delta(w_{d,s,i} = v, z_{d,s,i} = k)$ を除いた出現回数である．また，$n_{k,\cdot}$ は潜在トピック k の総割り当て数を表す．

式 (2) を定性的に解釈すると，右辺第 1 項目は文書中で頻出しやすいトピックへの割り当て確率が高くなることを示している．右辺第 2 項目は，その単語が頻出するトピックへの割り当て確率が高くなることを示

表 1: 提案モデルで用いられるパラメータ

D: 文書数
S: 文の数
K: トピック数
$n_{d,k}$: 文書 d でトピック k が表れた回数
$n_{k,v}$: 文書集合全体で語彙 v にトピック k が割り当てられた回数
$n_{d,s}$: 文書 d，文 s 中の単語数
$\boldsymbol{\alpha}$: Dirichlet 分布のパラメータ
$\boldsymbol{\beta}$: Dirichlet 分布のパラメータ
$\boldsymbol{\theta}_d$: 文書 d のトピック分布
$\boldsymbol{\phi}_k$: トピック k における単語の出現確率
$\boldsymbol{w}$: コーパス中の単語の集合
$\boldsymbol{z}$: トピックの集合
$w_{d,s,i}$: 文書 d，文 s 中の i 番目の単語
$z_{d,s,i}$: 文書 d, 文 s 中の i 番目の単語に割り当てられたトピック
$\boldsymbol{z}'_{d,s,i}$: $z_{d,s,i}$ に対応する要素を 1，他の要素を 0 にした one-hot ベクトル
$\bar{\boldsymbol{z}}_{d,s}$: 文書 d，文 s のトピックを単語数で正規化したベクトル
$\boldsymbol{e}_{d,s}$: 文書 d, 文 s の Sentence Embeddings
d_v: Sentence Embeddings の次元数
$\boldsymbol{\eta}^{\mathrm{T}}$: $\bar{\boldsymbol{z}}_{d,s}$ を正規分布の平均に変換する行列
σ^2: 正規分布の分散に関するパラメータ
$\boldsymbol{A}$: 各文のトピック分布を縦に並べた $S \times K$ の行列
$\boldsymbol{Y}$: 各文の分散表現を縦に並べた $S \times d_v$ の行列

している．一方式 (2) の右辺第 3 項目は，$\bar{\boldsymbol{z}}_{d,s}$ を行列 $\boldsymbol{\eta}^{\mathrm{T}}$ により変換したベクトルを平均とする正規分布からの Sentence Embeddings の生成確率が高くなるようにトピック割り当てが行われることを示している．この項を導入することにより，Sentence Embeddings を考慮したトピック割り当てが可能になる．

また，$\bar{\boldsymbol{z}}_{d,s}$ を変換する行列 $\boldsymbol{\eta}^{\mathrm{T}}$ は 1iteration 毎に以下の式を用いて更新される．

$$\boldsymbol{\eta} = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

4 評価実験

潜在トピックモデルにおける Sentence Embeddings の有効性を検証するために評価実験を行った．

4.1 実験設定

提案モデルの性能を評価するために，推定されたトピック分布を用いて文書分類を行い，LDA を用いた場合との精度の比較を行った．学習用･評価用データセットとして，20Newsgroups[1] を用いた．20Newsgroups は全 20 クラスのうちいずれかのラベルが付与された 18,846 文書から成り，訓練用データ (約 60%) と評価用データ (約 40%) に分割されている．今回の実験では訓練用データおよび評価用データとして，上記のデータをそのまま使用した．

前処理として，Stanford CoreNLP[2] により文分割，原形化を行った後，全ての文字の小文字化を行った．また，数字および記号文字，GloVe[3] に収録されていない単語をストップワードとして削除した．

Dirichlet 分布のパラメータである $\boldsymbol{\alpha}$ および $\boldsymbol{\beta}$ はそれぞれ 0.1, 0.01 とした．$\boldsymbol{\eta}^{\mathrm{T}}$ は全要素が 0.1 の行列で初期化を行い，σ^2 は 2.0 とした．トピック数 K は 20 とし，トピックモデル学習の際の反復回数は 500 回とした．提案モデルの実装は java で行った．[4]

文の分散表現化手法には以下の 3 つの手法を用いた．ただし，分散表現の次元数 d_v は 50 とした．

- Average Word Embeddings (AWE): 文中の各単語の分散表現の平均を Sentence Embeddings とする．単語の分散表現には GloVe により事前学習された 50 次元のベクトルを用いた．
- Paragraph2Vec (P2V) [4]: Paragraph2Vec により学習された分散表現を Sentence Embeddings とする．gensim[5] に実装されている doc2vec を使用し，分散表現を得た．
- Skip-Thought Vectors (STV) [5]: Skip-Thought Vectors により学習された分散表現を Sentence Embeddings とする．Chainer[6] を用いて実装を行った．

文書分類の際には，Support Vector Machine (SVM) により多クラス分類を行った (one-versus-rest 法を使用)．文書の特徴量として，各モデルにより推定されたトピック分布を用いた．実装には scikit-learn[7] を使用した．カーネルには RBF カーネルを用い，ハイパーパラメータ γ, C についてはデフォルトパラメータ ($\gamma = 0.1, C = 1.0$) を用いた．

[1] http://qwone.com/ jason/20Newsgroups/
[2] http://stanfordnlp.github.io/CoreNLP/
[3] http://nlp.stanford.edu/projects/glove/
[4] ソースコードは公開予定である．
[5] https://radimrehurek.com/gensim/index.html
[6] http://chainer.org/
[7] http://scikit-learn.org/

表 2: 文書分類の精度 (20 クラス分類)

Method	Accuracy (500 iteration)	Accuracy (Max)
LDA	0.47	0.52
Proposed (AWE)	0.55	0.56
Proposed (P2V)	0.54	0.55
Proposed (STV)	**0.56**	**0.57**

4.2 実験結果・考察

表 2 に文書分類の精度を示す．ただし，Accuracy (500 iteration) は反復回数が 500 回のときのモデルを用いた場合の精度である．Accuracy (Max) は各反復毎のモデル全てで精度を算出したうちの最高精度である．表 2 から，Sentence Embeddings を用いることで文書分類の精度が向上していることが分かる．文の分散表現化の手法を比較すると，Skip-Thought Vectors により得られた分散表現を用いた場合の精度が一番高いことが分かる．

図 2 に反復回数と分類精度の関係を示す．図 2 から，通常の LDA に比べて，Sentence Embeddings を用いた場合のモデルのほうが全体的に精度が高いことが分かる．また，反復回数が 80 回よりも小さい場合どのモデルも同じような分類精度であることが分かる．これは，反復回数が少ないうちは Sentence Embeddings の影響が小さいためであると考えられる．一方，反復回数が 80 回を超えたあたりから通常の LDA の精度は収束しているが，Sentence Embeddings を用いたモデルの精度は向上していることが分かる．これは，反復回数が多くなるにつれて，Sentence Embeddings の影響が大きくなったためであると考えられる．

表 3 に σ^2 の値を変更したときの文書分類の精度を示す．ただし，Sentence Embeddings には Average Word Embeddings (AWE) を用い，反復回数が 500 回のときのモデルを用いた．表 3 から，σ^2 の値により文書分類の精度が大きく変わることが分かる．今回は計算時間の関係で共分散行列を固定にしたが，$\boldsymbol{\eta}^{\mathrm{T}}$ と同様に更新することが理想的であると考えられる．この点は今後の課題である．

5 おわりに

本稿では，Sentence Embeddings を導入した潜在トピックモデルを提案し，その有効性について検証を行った．潜在トピックモデルにおいて Word Embeddings

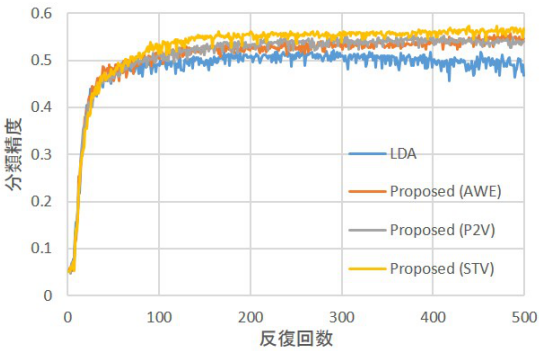

図 2: 反復回数と分類精度の関係

表 3: σ^2 と文書分類精度の関係

σ^2	Accuracy
0.1	0.32
1.0	0.51
2.0	0.55
3.0	0.56
4.0	0.52

を用いる研究は存在するが，Sentence Embeddings を導入しその有効性の検証を行ったのは本研究が初めてである．評価実験では，Sentence Embeddings を用いることで文書分類の精度が向上することを確認した．また，文の分散表現化手法についての比較も行い，Skip-Thought Vectors を用いた場合に分類精度が一番高くなることを確認した．

今後は，潜在トピックモデルにおいて Word Embeddings と Sentence Embeddings の同時適用を行う予定である．

参考文献

[1] David M Blei et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, pp. 993–1022, 2003.

[2] Rajarshi Das et al. Gaussian lda for topic models with word embeddings. *In Proceedings of ACL 2015*, pp. 795–804, 2015.

[3] Ximing Li et al. Integrating topic modeling with word embeddings by mixture of vmfs. *In Proceedings of COLING 2016*, pp. 151–160, 2016.

[4] Quoc Le et al. Distributed representations of sentences and documents. *In Proceedings of ICML 2014*, pp. 1–9, 2014.

[5] Ryan Kiros et al. Skip-thought vectors. *In Proceedings of NIPS 2015*, pp. 1–9, 2015.

[6] David M Blei et a. Supervised topic models. *In Proceedings of NIPS 2008*, pp. 1–8, 2008.