潜在情報に基づくテキストおよび脳活動データの要約

理学専攻・情報科学コース　尾崎 花奈 (指導教員：小林 一郎)

1　はじめに

近年，大量のテキストデータを効率的に処理して理解できるようにする技術の必要性が高まり，それに伴いトピック解析や文書要約などの技術が発展している．また，テキストデータだけに限らず，神経科学分野においてヒトの脳を解析する際にも大きな次元を持つ脳情報から重要となる情報を取り出して解析する技術が求められている．本研究では，テキストデータに対して効率的なアルゴリズムを採用したトピック抽出モデルの提案，要約文生成にトピック情報を加えたモデルの提案を行い，加えて，脳活動データに対して必要な情報を抽出して意味情報との対応関係の調査を行なった．

2　トピック抽出モデル

2.1　トピックモデル

トピックモデルは，文書の中に潜在的に存在するトピックを自動で抽出するモデルである．代表的な手法である LDA (Latent Dirichlet Allocation)[1] は，各文書に潜在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潜在トピックという非観測確率変数で定式化する．Das らによって提案された Gaussian LDA[2] は，LDA に単語の分散表現 (Word embedding) を組み合わせている．Gaussian LDA は単語の意味的関係性を事前知識として持つため，トピック内の意味的結束性が向上している．Gaussian LDA のグラフィカルモデルを図 1 に示す．

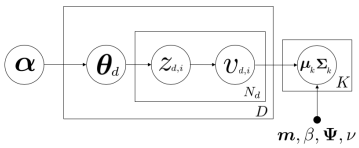


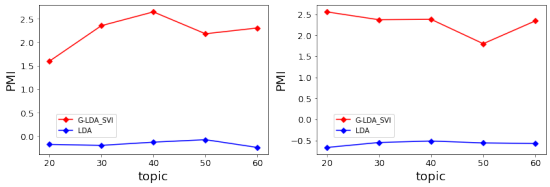
図 1: Gaussian LDA のグラフィカルモデル

2.2　SVI を用いたトピック推定

Gaussian LDA における事後分布推定では周辺化ギブスサンプリングを用いているが，本研究では SVI (Stochastic Variational Inference)[3] を用いることで，計算時間の短縮を実現している．ギブスサンプリングにおいては文書全体に対して繰り返し学習が必要であったが，SVI は文書を逐次的に学習する．

2.3　実験

本研究では，提案手法が従来の LDA に比べて単語の意味的関係性を捉えたトピック抽出をしているかを評価する実験を行なった．データセットとして，それぞれ 18846 文書，1740 文書から成る 20Newsgroups と NIPS の 2 つを用いた．ベクトル化された単語のデータとして，Wikipedia で Word2Vec により学習された 50 次元のデータを用い，トピック数 K は 20 から 60 まで 10 ごとに設定した．評価指標として，先行研究である Gaussian LDA が用いていた自己相互情報量 (PMI) を用いる．今回は各トピック上位 10 単語の PMI で評価した．図 2 は，2 つのデータそれぞれについて 20 から 60 の各トピックにおける提案手法と従来の LDA の PMI を表している．どのトピック数においても従来の LDA よりも PMI の値が上回っていることがわかる．



(a) 20Newsgroups　(b) NIPS

図 2: 20Newsgroup と NIPS における従来の LDA との各トピック PMI 比較

3　文書要約モデル

3.1　事前学習言語モデル

事前学習言語モデルは，大量のコーパスに対して文脈を学習させたモデルであり，従来の自然言語モデルと異なり 1 つのモデルを転移学習することで，文章分類、翻訳など様々なタスクへの応用が可能である．その中でも，Google が開発した BERT[4] は，様々な言語処理のタスクにおいて革新的な結果を達成している．

3.2　事前学習言語モデルを用いた要約モデル

文書要約は，自然言語処理におけるタスクの 1 つであり，テキストの自動要約を行う．要約の種類には抽出型と生成型があり，近年はニューラルネットを用いたモデルの発展により，生成型要約の研究が盛んに行われている．その中でも，事前学習言語モデルを採用したモデル BERTSUM[5] が注目されている．BERTSUM は，事前学習モデル BERT を拡張した文書レベルの Encoder を用いることによって，事前学習で得た言語モデルを文抽出や文生成に応用することに成功している．このモデルは抽出型要約と生成型要約のどちらにも対応しており，抽出型要約においては BERT が出力した文ベクトルから，その文が要約文に含まれるべきかどうかを学習し，生成型要約においては Decoder 部分に Transformer を採用して BERT が出力したトークンベクトルを入力として要約文を生成している．

3.3　トピックの導入

本研究では，BERTSUM における BERT の出力にトピックベクトルを加えることで，文書中のトピックを考慮した要約モデルを提案する．訓練文書に対して LDA によるトピック解析を行い，得られた 2 つの分布 (文書ごとのトピック分布，トピックごとの単語分布) のアダマール積をとることによって，トークンごとのトピックベクトルを得る．抽出型要約においては BERT が出力した文ベクトルに対して，トークンごとのトピックベクトルを文ごとに加算したベクトルを結合し，以降の処理の入力とした．生成型要約においては BERT が出力したトークンベクトルにトークンごとのトピックベクトルを結合し，Decoder の入力とした．

3.4 実験

本研究では，トピックベクトルを BERTSUM に加えることで要約の精度が上がるかどうかを評価する実験を行なった．また，本稿では，extractive summarization について，summarization layer に Transformer を採用した結果を掲載した．データセットとして CNN/Daily Mail を用いた．CNN/Daily Mail データセットはニュースの記事とそのハイライトから成るデータであり，訓練，検証，評価用の割合は，287,227/13,368/11,490 文書とした．トピック数は 512 とした．テスト評価指標としては，正解要約文と一致する単語の割合を数える ROUGE を用いた．結果を表 1 に示す．

	R1	R2	R3
BERTSUM	42.93	20.11	39.37
提案手法	42.76	20.01	39.20

表 1: CNN/Daily Mail データセットにおける ROUGE 値

BERTSUM に比べて提案手法の方が ROUGE の値が下がっている．これは，トピック抽出の際にストップワードを除かなかったこと，トピック数が多すぎたことが要因だと考えられる．

4 脳活動データ分析

4.1 脳活動データから意味表象の推定

近年，動画像などを視聴した際の脳の活動パターンから人がどのような意味カテゴリを想起しているかを調査する研究が盛んになっている．本研究では，動画視聴時のヒトの脳活動と，その動画の説明文が対応した行列に対して辞書学習を行うことで獲得した辞書基底を利用して，テスト用脳活動データに対して脳活動に対応する動画像説明文のベクトルの推定を行なった．

4.2 実験

使用するデータは，動画視聴時の 3 人分の脳活動データと動画説明文である．脳活動データは，被験者に動画を見せ，fMRI を用いて脳神経活動を記録したものである．また，実験に使用するボクセルは先行研究 [6] で予測精度がある閾値以上になったボクセルとした．脳活動データの各ボクセルの観測値を入れて行列化して．脳活動行列を作成した．動画説明文は動画像から 1 秒ごとに抽出した静止画に対し，アノテータ 40 人からランダムに抽出された 4 人が静止画を見て想起したことを文章にしている．動画説明文からサンプルごとに出現する単語 (名詞，動詞，形容詞) の分散意味表現の平均ベクトルからなる行列を作り，これを意味表象行列とする．これら 2 つの行列を縦に結合させ，脳活動と意味表象の結合行列を作成する．被験者が動画を見てから fMRI で観測される脳活動との時間差を考慮し，脳活動行列と意味表象行列を 4 秒または 6 秒ずらして結合した．訓練では，結合行列に対して辞書学習を行う．テストでは，同じ方法で作られた脳活動行列と，訓練で得られた辞書行列を用いてスパースコーディングを行い，得られた行列を推定意味表象行列とする．提案手法は図 3 に示した通りである．

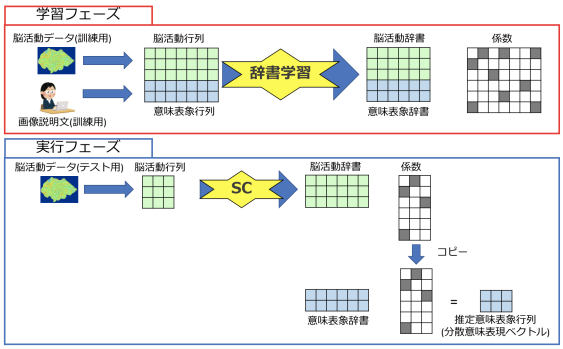


図 3: 意味表象推定方法

4.3 結果

表 2 は，正解意味表象行列と推定意味表象行列の cos 類似度のマクロ平均を表しており，それぞれ観測の時間差ごとに示している．

被験者	基底数	時間差	
		4 秒	6 秒
A	800	0.359	0.379
B	900	0.933	0.932
C	1100	0.235	0.226

表 2: 正解意味表象と推定意味表象との cos 類似度

5 おわりに

本研究では，テキストから単語の意味的関係性を捉えたトピック抽出を行うモデルに高速化アルゴリズムを採用し，実験にてトピック内の意味的一貫性が向上していることと実行時間の高速化を確認した．また，トピックモデルをテキスト要約モデルに応用し，文書内のトピック情報を捉えた要約モデルを構築し，実験を行なった．さらに，ヒトの動画視聴時の脳活動データと意味表象のペアデータを辞書学習することにより，これらの対応関係を捉えた基底を作ることに成功した．

今後の課題としては，文書内のトピックを利用して生成する要約文を制御可能なモデルへ発展させることや，脳活動データと意味表象の対応関係を詳しい調査が上げられる．

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pp. 601–608. MIT Press, 2002.

[2] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *ACL (1)*, pp. 795–804. The Association for Computer Linguistics, 2015.

[3] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, Vol. 14, No. 1, pp. 1303–1347, May 2013.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Yang Liu. Fine-tune BERT for extractive summarization. *CoRR*, Vol. abs/1903.10318, , 2019.

[6] Satoshi Nishida, Alexander G. Huth, Jack L. Gallant, and Shinji Nishimoto. Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. *Society Neuroscience Abstract*, Vol. 45, p. 333.13, 2015.