事前学習モデルと潜在トピックを用いた文書要約への取り組み

尾崎 花奈 †
† お茶の水女子大学 人間文化創成科学研究科
理学専攻 情報科学コース

小林 一郎 ‡
‡ お茶の水女子大学
基幹研究院 自然科学系

{ozaki.kana, koba}@is.ocha.ac.jp

1 はじめに

大量のテキストデータが存在する現在，文書の概要を生成する文書要約技術の必要性が高まっている．要約の手法には抽出型と生成型があり，抽出型は原文書の表現をそのまま使って要約文を構築する手法であり，生成型は原文書にないフレーズや単語も含めて生成的に要約文を構築する．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モデル BERT[1] を採用した Bertsum[2][3] は，抽出型要約と生成型要約のどちらにも対応し，どちらの手法においてもベースラインに比べて要約の精度が上がったと報告している．一方で，近年，要約においてトピック情報を用いることによって精度が向上したという研究が報告されている [4]．本研究では，Bertsum 対して，文書内のトピック情報を加えたモデルを提案し，抽出型要約において実験を行なった結果を報告する．

2 関連研究

Liu らによって提案された Bertsum[2][3] は，事前学習モデル BERT[1] 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．事前学習言語モデルとは，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 が開発した自然言語処理深層学習モデルである BERT 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．Bertsum 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ている．分類を行う summarization layer には Simple Classiffier，RNN，Transformer をの 3 種類を採用して実験し，Transformer を用いたモデルが最も評価値 (ROUGE) が良かったと報告している．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．

Narayan ら [4] は CNN ベースの seq2seq モデルである ConvS2S[5] に対して入力文書のトピック情報を補助的な情報としてモデルに追加している．結果として Encoder 側，Decoder 側にもトピックベクトルを加えた T-ConvS2S が比較手法の Pointer Generator Network を採用したモデルに比べて生成型要約における ROUGE の値が向上したと報告している．

3 提案手法

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[6] 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．トピックベクトルの作成にあたっては，Narayan らによるトピックベクトルを採用した CNN ベースの生成型要約の研究 [4] に倣った．2 つの分布のうち文書ごとのトピック分布は，対象単語が含まれている文書についての分布を取ってくる．トピック数を K とし，$t_D \in \mathbb{R}^K$ を文書 D のトピック分布，$\mathbf{t}' = (t'_1, ..., t'_m)$ を文書内の単語ごとのトピック分布とする．ここで，$t'_i \in \mathbb{R}^K$ とする．これにより単語ごとのトピックベクトル c_i は，

$$c_i = t'_i \otimes t_d \in \mathbb{R}^k, \tag{1}$$

で表され，$\otimes$ は要素ごとの積 (アダマール積) を表す．単語 w_i のトピック分布 t'_i は本質的にその単語自体の

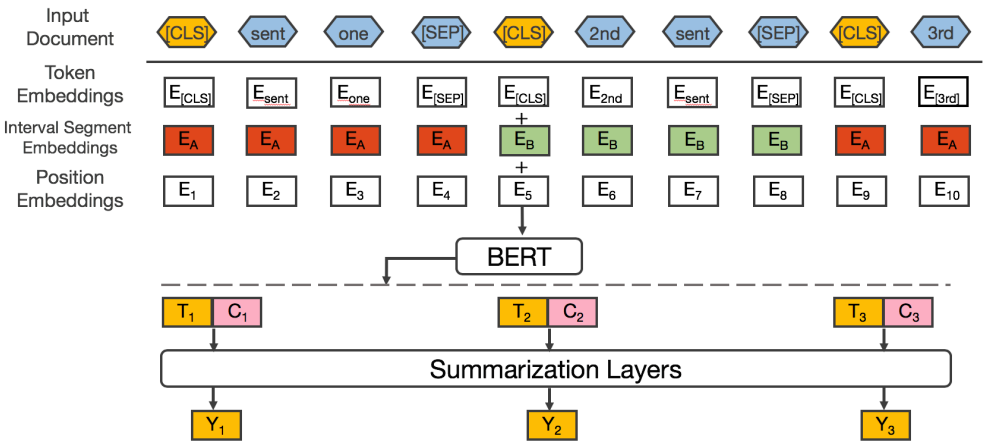

図 1: 提案手法概要図

局所的な特徴を捉えており，文書ごとのトピック分布 t_d は文書全体の特徴を表している．この 2 つのベクトルのアダマール積を BERT の出力ベクトルに加えることで，文脈の情報からベクトル化されたトークンに対してトピックの情報を追加したものとなる．

本稿では，抽出型要約モデルに対してトピックベクトルを導入したモデルでの実験を掲載する．抽出型手法においては文ごとのベクトルを使用するので，トークンごとにつけられたトピックベクトルを文ごとに加算し，それを BERT が出力した文ベクトルに結合し，以降の入力とした．モデルの概要図を図 1 に示す．

4 実験

本稿では，トピックベクトルを BERTSUM に加えることで抽出型手法における要約の精度向上につながるかを検証する実験を行なった．BERTSUM のモデル構造に従って Liu ら [2] の評価実験の設定と同じく summarization layer を Simple Classificaton，RNN，Transformer の 3 種類で実験した．

4.1 データ

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記事とそのハイライトが含まれているデータであり，文書要約の研究においてはハイライトを正解要約文として使用している例が多く見られる．訓練，検証，評価用データの割合は，282,227/13,368/11,490 文書とした．

CNN/Daily Mail データの詳細を表 1 にまとめた．

データの前処理として，Stanford CoreNLP を用いて文分割を行い，先行研究 [2][7] に従って処理を行い，入力文書は全て 512 単語に切り捨てた．また，入力文書は 3 文以上 100 文以下とし，入力文書の各文については 5 単語以上 200 単語以下とした．要約文は 5 単語以上 500 単語以下とした．

4.2 トピック数の決定

LDA におけるトピック数の決定に際し，Perplexity と Coherence を評価指標としてグリッドサーチを行なった．トピック数は {32, 64, 96, 128, 160, 192, 224, 256, 512, 768} に設定し，各設定において訓練文書に対する Perplexity と Coherence を求めた．モデルの学習にはライブラリ gensim[1]に実装されている Stochastic Variational Inference を用い，事前処理としてストップワードを除去した．Coherence は Mimno ら [8] による学習コーパスのみで Coherence を算出する手法を用いた (ライブラリ gensim の UMass Coherence として実装されている)．結果は図 2 のようになった．

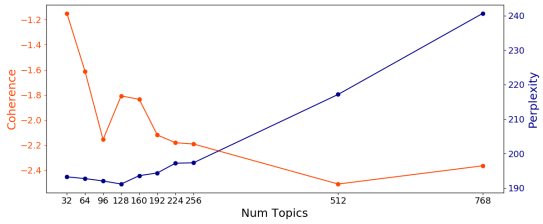

図 2: トピック数による Perplexity と Coherence

[1] https://radimrehurek.com/gensim/

データセット	文書数 (訓練/検証/評価)	文書 (記事)		要約文 (ハイライト)	
		単語数	文数	単語数	文数
CNN	90,266/1,220/1,093	760.70	33.98	45.70	3.59
Daily Mail	196,961/12,148/10,397	653.33	29.33	54.65	3.86

表 1: CNN/Daily Mail データセット詳細

Perplexity は低いほどよく，Coherence は高いほど良いモデルであるとされている．図 2 を見ると，Perplexity はトピック数 128 で収束したあと上昇し，Coherence は最初こそ高いものの，128 を極大値として徐々に低くなっていることがわかる．これらの結果からトピック数としては 128 を採用した．表 2 にトピック数を 128 に設定した際に LDA によって抽出されたトピックの例を示す．

T1:	hotel, mar, resort, spa, castro, top, hotels, list, california
T2:	eu, migrants, italy, island, lamp, sea, europe, italian, lindsay, european
T3:	mrs, patients, hospital, said, doctors, medical, treatmet, care, health
T4:	club, league, football, team, players, season, fans, cup, last, liverpool, game, england, match
T5:	space, earth, pilot, pip, brit, station, moon, samantha, astronauts, esa, engineer

表 2: CNN/Daily Mail データセットにおいて LDA によって抽出されたトピック (訓練文書)

4.3 実験設定

実行環境として GPU(Tesla V100) を 3 枚用い，実行時間は約 16 時間であった．抽出する要約文はモデルの出力スコアが高かった順に 3 文とした．また，先行研究に従い，最終的な要約文抽出過程において要約文の冗長性回避のため，Trigram Blocking を採用した．すでに選ばれた要約文 S と新たに加えたい要約文 c に対して，c と S の間に trigram 以上の重複があった場合には c を要約文から除いている．実験の詳細を表 3 にまとめた．

また，CNN/Daily Mail データセットにおける記事のハイライトは文抽出により作られた文ではないため，抽出型要約の訓練用として原文書の各文が要約に含まれるべき文かどうかを示すアノテーションが必要である．本研究では BERTSUM で実装されているアルゴリズムに従い，ハイライトを正解要約文とした時の ROUGE-2 F1 スコアを最大にする文集合 (ORACLE) を抽出するべき文集合とした．

BERT 出力次元数	768 次元
トピック数	128
隠れ層次元数	896 次元
訓練ステップ数	50000 ステップ
最適化手法	Adagrad
誤差関数	交差エントロピー
バッチサイズ	3000
チェックポイント	1000 ステップ

表 3: 実験設定詳細

4.4 評価指標

評価には，一般的に要約評価に用いられている ROUGE score を採用した．ROUGE は人間が作った要約 (正解) とシステムが作った要約の一致度を測る指標であり，ROUGE-N は N-gram 単位での一致を取り，ROUGE-L は一致する最大シーケンスを評価する．先行研究に従い，検証データセットにおいて loss が低かった上位 3 つのチェックポイントにおけるテストデータセットでの ROUGE-1/ROUGE-2/ROUGE-L (F1 値) の平均を求めた．

4.5 結果

CNN/Daily Mail データセットにて実験を行なった結果を表 4 に示す．まず，CNN/Daily Mail データセットにおける文抽出 ROUGE 値の最大値は ORACLE システムでの ROUGE 値である．また，ベースラインとし

モデル	R1	R2	RL
Oracle	52.59	31.24	48.87
Lead-3	40.42	17.62	36.67
Transformer	40.82	18.18	37.15
Bertsum+Classifier	42.87	20.05	39.30
Bertsum+RNN	42.91	20.06	39.34
Bertsum+Transformer	**42.93**	**20.11**	**39.37**
T-Transformer	40.57	18.06	36.93
T-Bertsum+Classifier	42.85	20.05	39.27
T-Bertsum+RNN	42.82	20.04	39.26
T-Bertsum+Transformer	42.82	20.03	39.27

表 4: CNN/Daily Mail データセットにおける ROUGE 値.

て示した Lead-3 は文書から最初の 3 文を取って来たものである．比較手法として本研究でベースのモデルとして使用した Bertsum について，summarization layer が Classifier，RNN，Transformer の 3 種類と Liu らにより実装されていた事前学習なし (パラメータをランダムに初期化して要約タスクでのみ学習) の Transformer モデルについて結果を掲載した．Bertsum については，本研究と同じ実行環境で再現実験を行なった結果である．提案手法の結果は表の一番下に示した T を冠する表記のものであり，Bertsum にトピックベクトルを追加した 4 種類である．

4.6 考察

先行研究である Bertsum に比べて 4 種類のモデル全てにおいてわずかに ROUGE スコアが下がっており，トピックベクトル導入による要約の精度の向上は見られなかった．しかし，言語モデルに由来する ROUGE スコアでは，入力文書のトピックを捉えた要約文を正しく抽出しているかを測ることはできないと考えられる．また，入力文書からそのまま文を抽出する抽出型要約においてはトピックを反映した文生成による要約ではないため，トピックを入れたことによる効果は現れにくかったと考えられる．

5 おわりに

本研究では，事前学習を用いた要約モデルを基にして，入力文書のトピック情報を追加したモデルを提案した．評価実験においてトピック情報を追加した際の ROUGE スコアによる要約精度の向上を確認することはできなかった．しかし，ROUGE スコアは要約文のトピックを測る指標ではないため，提案手法に即した要約文評価の指標が必要であると考えられる．

今後の課題として，正解要約文のトピックとモデルが抽出した要約文のトピックの一致を測る指標を取り入れたい．また，要約文にトピックが反映されやすいと考えられる生成型要約においても提案手法の有効性を検証するつもりである．

参考文献

[1] 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*.

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

[3] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders, 2019.

[4] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[5] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *CoRR*, Vol. abs/1705.03122, , 2017.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, March 2003.

[7] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

[8] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.