トピックに基づく抽象型要約文生成への取り組み

横川　悠香　(指導教員：小林 一郎)

1　はじめに

大量の情報が溢れている現代において，情報を短時間で把握できる要約文の生成を行う研究が盛んに行われている．要約には大きく分けて 2 種類の手法がある．1 つは，要約元の文章から重要文を抽出することにより要約文を生成する「抽出型要約」であり，もう 1 つは人が行うように文書内容を反映した要約文を生成する「抽象型要約」である．深層学習が要約文生成研究に導入された後は，後者の研究が盛んに進められている．本研究でも抽象型要約を行い，文章生成のタスクにおいて用いられる，文章の持つスタイルをコントロールする手法を要約文の生成に取り入れた手法を提案する．

2　提案手法

図 1 に提案手法の概要を示す．

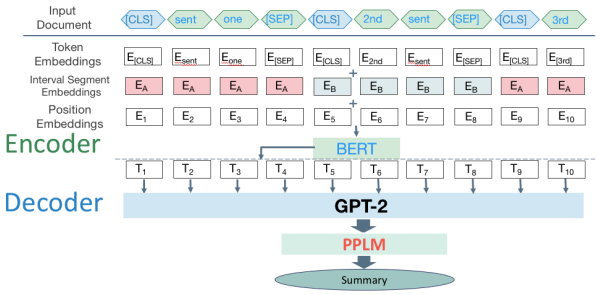

図 1: 提案手法

抽象型要約を行うエンコーダ・デコーダ手法である PreSumm [1] をベースとし，デコーダを GPT-2 [2] に変更する．変更したモデルが要約文を生成する際に，PPLM [3] の手法を取り入れることによって，要約文にトピックを導入する．

2.1　PreSumm

PreSumm [1] は，Transformer [4] ベースのエンコーダ･デコーダ要約手法である．エンコーダには BERT [5] を用いており，通常の BERT の単語埋め込みに変更を加えることを特徴としている．要約対象の文をエンコーダである BERT に与え，得られた隠れ状態をデコーダである Transformer に渡し，要約を生成する．

2.2　GPT-2

Generative Pre-Training 2（GPT-2）[2] は，Transformer のデコーダ部分のみを用いて開発された汎用言語モデルであり，特に文章生成タスクにおいて高い性能を示している．GPT-2 は，BERT のように大量のデータを使って特定のタスクに適するようにファインチューニングする必要がない，教師なし学習で様々なタスクを学習可能であることを目的として作られた汎用言語モデルである．言語モデルにおいて，タスクを含めて出力を考え，大量のデータを用いてモデルを学習するため，一度も見たことがない（zero-shot setting）データに対しても精度良く推定が行える．

2.3　PPLM

Plug and Play Language Models（PPLM） [3] は，文章生成においてトピックや感情を文に導入する手法である．その特徴として，既存の言語モデルに追加学習することなく，文にトピックを取り入れられる点が挙げられる．

2.3.1　PPLM の着想

いま，x を生成されたサンプル，a をコントロールしたい分布とする．文のスタイルをコントロールするということは，$p(x|a)$ をモデリングすることを含意する．しかし，言語モデルは $p(x)$ のみを学習している．そこで、ベイズの定理により

$$p(x|a) \propto p(x)p(a|x) \tag{1}$$

であることから，$p(x)$ に分布 $p(a|x)$ を掛け合わせることにより，$p(x|a)$ を得る．PPLM では，$p(a|x)$ にトピックを導入する場合は Bag of Words，感傷的な文といった感情の導入を行う場合は単層の識別器を用いる．本研究では前者を用いたトピックの導入を要約文生成に加える．

2.3.2　PPLM によるトピックの導入

トークンの列 $X = \{x_0, \ldots, x_n\}$ が与えられたとき，言語モデル（LM）は $p(X)$ を学習する．H_t を Transformer のアテンション計算における key-value のペアで構成される行列とする．

$$H_t = [(K_t{}^{(1)}, V_t{}^{(1)}), \ldots, (K_t{}^{(l)}, V_t{}^{(l)})] \tag{2}$$

ここで，$(K_t{}^{(i)}, V_t{}^{(i)})$ は i 番目の層の，0 から t までの全てのタイムステップで生成された key-value のペアである．PPLM では，2.3.1 節で挙げた着想をもとに，生成される文がコントロールしたい分布 a を持つように式 (2) における H_t をアップデートする．ΔH_t を H_t に対するアップデートとする．分布 $p(a|x)$ を $p(a|H_t + \Delta H_t)$ と置き換えた上で ΔH_t は以下のように計算される．

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{||\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)||^{\gamma}} \tag{3}$$

Wolf らによる効率的な Transformer の実装 [6] に基づき，アップデートされた $\tilde{H}_t = H_t + \Delta H_t$ を用いて，x_t が与えられたときの $\tilde{x}_{t+1}$ が以下のように予測される．

$$\tilde{o}_{t+1}, H_{t+1} = LM(x_t, \tilde{H}_t) \tag{4}$$

$$\tilde{x}_{t+1} \sim \tilde{p}_{t+1} = Softmax(W\tilde{o}_{t+1}) \tag{5}$$

$\tilde{o}_{t+1}$ はロジットベクトルであり，W は線形変換である．このように，コントロールしたい分布に基づいてサンプルが生成される．

本研究では，入力をエンコーダに通して得られた隠れ状態をもとに，デコーダで要約文を生成する段階において PPLM のトピック導入を追加し，要約文にトピックを付加する．

3 実験

3.1 実験設定

CNN/Daily Mail Dataset [1]を使用する．このデータセットはニュース記事とその短い要約からなり，文章要約の先行研究において広く使用されている．データ数は訓練/検証/評価で 287227/ 13368/11490 である．データの前処理は PreSumm と同様に行なう．エンコーダの入力である記事は最大 512 単語に切り捨てる．要約の正解文のトークナイザーは GPT-2 のものに変更した．

表 1: 実験設定

データセット	CNN/Daily Mail
学習ステップ	200000
勾配法	Adam
語彙	エンコーダ:30522 デコーダ:50259
損失関数	負の対数尤度
学習率	エンコーダ:0.002 デコーダ:0.2
BoW の単語数	Positive:2005 Negative:206 Legal:131

トピックの導入に用いる Bag of Words は，ポジティブな単語，ネガティブな単語，法律に関わる単語の 3 種を用いる．ポジティブな単語と法律に関わる単語は PPLM で使用される単語リストを使用する．ネガティブな単語は 4600 個の単語リスト[2]の中から，CNN/Daily Mail Dataset での出現回数が上位 200 個から 3200 個に入る単語のみを使用する．

評価指標には ROUGE，Distinct-N を用いる．ROUGE [7] は文章要約の評価指標として多く使用されるもので，PreSumm においても用いられている．Distinct-N [8] は文章の多様性を評価する指標である．Distinct-N の計算は各モデルの出力から最初の 1000 文書を抜き出して行った．

3.2 実験結果

表 2: ROUGE スコア

モデル	R1	R2	RL
BertSumAbs	41.72	19.39	38.76
GPT-2	37.12	15.74	34.63
GPT-2 + PPLM			
(Positive)	37.07	15.71	34.59
(Negative)	37.04	15.65	34.55
(Legal)	37.01	15.65	34.54

表 3: Distinct-N スコア

モデル	Dist-1	Dist-2	Dist-3
GPT-2	0.8737	0.9486	0.9234
GPT-2 + PPLM			
(Positive)	0.8737	0.9472	0.9220
(Negative)	0.8745	0.9491	0.9233
(Legal)	0.8760	0.9472	0.9211

3.3 考察

表 2 では，本研究での PPLM を加えないモデルにおいて，PreSumm のモデルから ROUGE スコアが下がった．これは，デコーダを Transformer から GPT-2 にしたことにより，GPT-2 の言語モデルが文生成に大きく関与したことが原因ではないかと推察される．

また，PPLM を導入しない場合とする場合の ROUGE スコアの変化は小さかった．これは，ROUGE は要約の精度を測るために正解要約文との一致度を見る指標であり，トピックの追加は必ずしも精度の上昇には繋がらないとも考えられる．

表 3 では，ポジティブな単語の導入において PPLM を加えないモデルから Distinct-N のスコアが下がる結果となったが，ネガティブな単語の導入においては Dist-1，2 でわずかながらスコアの向上が見られた．また，ネガティブなトピックを導入して生成された要約文において，事件などのネガティブな記事の要約ほど PPLM によって導入された単語が出現しやすいことが確認できた．反対に，ポジティブな記事の要約にはあまり変化が見られなかった．ポジティブな場合も同様であり，PPLM によるトピック導入は導入したいトピックと要約する対象の記事が一致しているとその影響が大きくなると考えられる．このことから，法律に関するニュースは数が限られるため，法律に関する単語の導入の効果は限定的になったことも考えられる．

4 まとめ

本研究では，既存の抽象型要約手法である PreSumm のデコーダに汎用言語モデル GPT-2 を採用し，PPLM によるトピックを追加する要約文生成手法を提案した．CNN/Daily Mail Dataset を用いた実験から，一部においてはトピックの導入が確認されたが再検討の余地が残った．今後は，提案手法をさらに発展・応用し，既存研究との性能比較をしていきたい．

参考文献

[1] Yang Liu et al. Text summarization with pretrained encoders. In *Proc. of EMNLP-IJCNLP2019*, pp. 3730–3740, 2019.

[2] Vanya Cohen and Aaron Gokaslan. Opengpt-2: Open language models and implications of generated text. *XRDS*, Vol. 27, No. 1, p. 26–30, sep 2020.

[3] Sumanth Dathathri et al. Plug and play language models: A simple approach to controlled text generation. In *Proc. of ICLR2020*, 2020.

[4] Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[5] Jacob Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT2019*, pp. 4171–4186, 2019.

[6] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP2020*, pp. 38–45, 2020.

[7] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[8] Jiwei Li et al. A diversity-promoting objective function for neural conversation models. In *Proce. of NAACL-HLT2016*, pp. 110–119, San Diego, California, 2016.

[1]https://cs.nyu.edu/ kcho/DMQA/

[2]https://positivewordsresearch.com/list-of-negative-words/