Gated Convolutional Neural Network を用いた生成モデルの提案

1　はじめに

タイトルは文書を簡潔に表したもので，元文書の内容理解を容易にする．優れたタイトルは多くの人に元文書を読む機会を与えるものであり，そのようなタイトル生成は重要である．またインターネットの普及によりタイトル生成の需要は高まっている．しかしインターネット上のデータは指数関数的に増加しており，人手によるタイトル生成はコストが非常にかかるようになってきている．そこで，近年この問題点を解決する文書要約に関する研究が盛んに行われている．そのなかでもニューラルネットワークを用いた手法が注目を集めており，機械翻訳や文書要約などで高い精度が報告されている．

本研究では，入力単語に荷重をかけながら文章を生成していく Attention モデルに Gated convolutional neural network (GCNN) を用いたタイトル生成モデルを提案する．GCNN を用いることで，n-gram による単語間の特徴が得られるようにした．また GCNN の組み込みと GCNN の多層スタックを用いることで，Attention に用いる単語を削減するような GCNN 組み込み多層スタックモデルを提案する．

GCNN の組み込みにより，n-gram 間の単語の特徴が得られたかどうかを ABS モデルと ROUGE 値で比較し，提案モデルの性能を評価する．

2　要素技術

2.1　Attention モデル

機械翻訳などのタスクで考案された Encoder-Decoder モデルは可変長の文を固定長のベクトルにエンコードするため，長い入力文になるほど隠れ層のノード数が不足し，学習が難しくなる問題がある．そこで Bahdanau らは 2015 年，エンコーダ側で入力文の各単語の荷重を決定してエンコードするべき場所を制御する Attention モデルを提案した [1].

Attention モデルでは入力文の各単語 x_i に対する荷重 α_t を計算することで，エンコーダで出力される中間表現 c_t を得る．α_t はエンコーダで出力されるベクトル h_i とデコーダから出力されるベクトル h_t の内積を正規化することによって得られる．

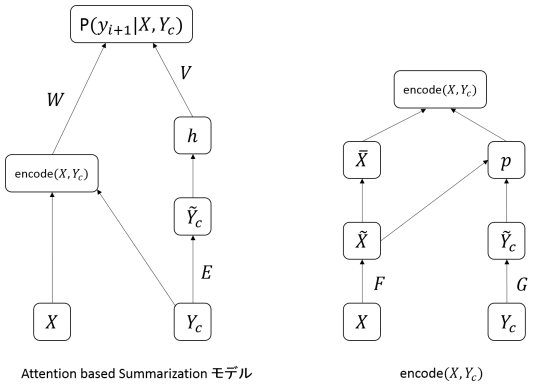

図 1: ABS モデル (文献 [2] をもとに作成)

$$\alpha_t(i) = \frac{\exp((\overline{h_i}, h_t))}{\sum_{j=1}^{n} \exp((\overline{h_j}, h_t))} \quad (1)$$

$$c_t = \sum_{i=1}^{n} \alpha_t(i)\overline{h_i} \quad (2)$$

2.2　Attention based Summarization モデル

Rush らは 2015 年，機械翻訳モデルである Attention モデルをもとに，文から短い文への生成型要約を行う Attention Based Summarization（ABS）モデルを提案した [2]．このモデルは Attention モデルの持つ単語に荷重をかけながらエンコードするという特徴を利用することで，要約に必要な単語に荷重をかけながら，生成的に文を要約することを可能とした．Rush らはこのモデルでニュース記事からタイトル生成をする実験により，従来の手法より高い精度を得た．図 1 に ABS モデルを示す．

ABS モデルはエンコーダ部分とデコーダ部分に分けられる．エンコーダ部分は 2.1 節で述べた Attention モデルが使用され，デコーダ部分には言語モデルが用いられる．

図 1 において X は元文を表し，Y_c は生成した c 個の単語を表す．E，F，G は元文，生成文の単語を 1-hot ベクトルから実数値ベクトルに変換するエンベディング行列であり W，V は重みパラメータであ

る．また p は単語の荷重，$\overline{X}$ は $\tilde{X}$ をスムージングした行列である．

それぞれのモデル式は以下の通りである．q はスムージングウィンドウサイズの大きさを，cat 関数はベクトルを連結する関数を表す．

$$\mathrm{encode}(X, Y_c) = p^{\mathrm{T}}\overline{X} \tag{3}$$

$$p \propto \exp(\tilde{X} P \tilde{Y}_c) \tag{4}$$

$$\overline{x}_i = \frac{\sum_{j=i-q}^{i+q} \tilde{x}_j}{q} \in \overline{X} \tag{5}$$

$$\tilde{x}_i = F x_i \in \tilde{X} \tag{6}$$

$$\tilde{Y}_c = \mathrm{cat}(G y_{i-c+1}, ..., G y_i) \tag{7}$$

$$\mathrm{P}(y_{i+1}|X, Y_c) \propto \exp(Vh + W\mathrm{encode}(X, Y_c)) \tag{8}$$

2.3 Gated Convolutional Neural Network

Gated Convolutional Neural Network (GCNN) は畳み込み層，Gated Linear Unit (GLU) 層のブロックを持つモデルであり，ゲート構造により，有用な情報の取捨選択が出来る [3]．畳み込み層で長期依存を捉え，GLU 層で上層に送る情報を制御する構造になっている．最終的にブロックの入力から出力を残差接続する．畳み込み層のカーネルサイズ k に応じて入力位置 k 個分の情報を集約できる．

3 提案手法

ABS モデルはエンコーダ側で単語間における特徴をとる際に単純なスムージングのみを用いており，文生成の際に単語のつながりを十分に考慮できていないモデルである．そこで本実験では，エンコーダ側に GCNN を用いることで単語間のつながりを考慮した n-gram 特徴量を得られるモデルを提案する．また GCNN を用いることで Attention の計算に用いるベクトルを削減することができる．

本提案モデルでは新たに 2 つのパラメータ *kernel_width*，*stack_size* を導入する．パラメータ *kernel_width* は何単語間の特徴を取るかを選択するパラメータで，パラメータ *stack_size* は GCNN を何層積むかを選択するパラメータである．

この 2 つのパラメータにより，Attention に用いる単語数を削減できる．削減単語数は以下の式で表される．

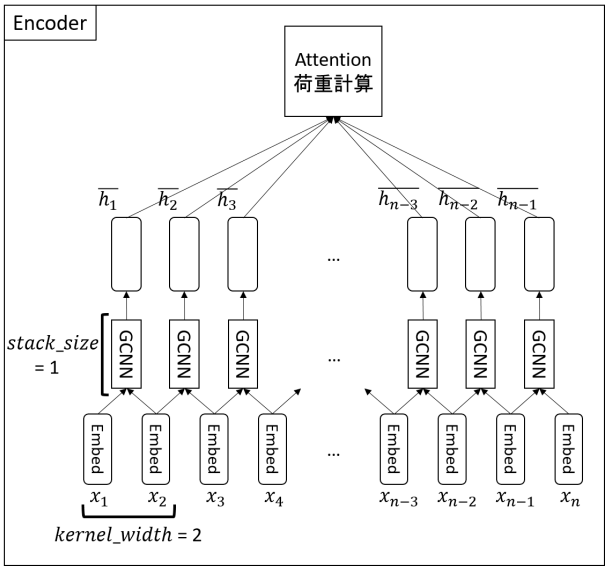

図 2: GCNN の内部構造

$$(kernel\_width - 1) \times stack\_size \tag{9}$$

一例として図 2 に $kernel\_width = 2$，$stack\_size = 1$ におけるモデルのエンコーダ部を示す．

4 実験

4.1 実験 1

4.1.1 実験方法

提案モデルでパラメータ *kernel_width*，*stack_size* を変えながらタイトル生成をし，ABS モデルと比較することで性能を評価した．

初めにデータセットについて説明する．毎日新聞データセット[1]の 2008 年から 2012 年までのデータのうち，社会，国際，経済に関するニュース記事を用いた．このうち 17000 件を学習データに，100 件をテストデータとして扱った．提案モデルにおいてエンコーダ側に用いる入力文としてニュース記事の最初の 1 文 (ヘッドライン) を用い，それに対してデコーダ側の出力文としてニュース記事のタイトルを用いた．ヘッドラインからタイトルを生成するように学習し，Validation 誤差が最小となった時点で学習を終了した．

なお形態素解析には Mecab を用い，数字に関してはすべて Num トークンに置換した．語彙数 20000

[1] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

表 1: 提案モデルにおけるパラメータ

パラメータ	値
エンベディングサイズ	200
隠れ層サイズ	200
kernel_width	2〜8
stack_size	1〜8
optimizer	Adam(alpha=0.001)
バッチサイズ	100

になるように単語の頻出数が少ないものから unk トークンに置換し，未知語として扱った．モデルは chainer フレームワークで実装した．モデルのパラメータは表 1 の通りである．

最後に文章生成方法と評価方法について説明する．実験では出力される要約文に対する負の対数尤度を最小化するように学習をしたため，要約文としては最も生成される確率の高い文章を選ぶ必要がある．これは語彙数 V の状態をもった最適経路問題を解くという問題に帰着されるが，これは NP 困難であり計算時間内に解くことが不可能である．そのため，枝刈りをしながら探索するために本研究では探索アルゴリズムとしてビームサーチを用いた．ビームサイズは 2 に設定した．

評価方法として，システム要約（生成文）と参照要約（正解文）間において n-gram 単位でどれほど一致したかを表す ROUGE [4] という指標を用いた．そのうち，ROUGE-N,ROUGE-S,ROUGE-L を用いた．今回最も着目した評価である ROUGE-N の計算式を以下に示す．

$$\text{ROUGE-N}(C,R) = \frac{\sum_{e \in \text{n-gram}(C)} \text{Count}_{\text{clip}}(e)}{\sum_{e \in \text{n-gram}(R)} \text{Count}(e)} \quad (10)$$

n-gram(C) は，システム要約に含まれる n-gram，n-gram(R) は，参照要約に含まれる n-gram 集合を表す．Count(e) は，ある n-gram の出現頻度を数える関数であり，$\text{Count}_{\text{clip}}(e)$ は，システム要約における n-gram の出現頻度 Count($e \in$ n-gram(C)) と参照要約における n-gram の出現頻度 Count($e \in$ n-gram(R)) の小さい方の値を採用する．

4.1.2 実験結果

初めに提案モデルに対するパラメータを変えた際のテストデータ 100 件の ROUGE 値を表 2 に示す．

表 2: t 検定

	t 値	p 値
ROUGE-1	0.5454	0.5454
ROUGE-2	0.5454	0.5454
ROUGE-3	0.5454	0.5454

また実際の生成タイトル例を表 3 に示す．

表 2 の結果から提案モデルにおいて *kernel_width* が 2,3 のとき，ABS モデルと比較して ROUGE-1 については差はあまり見られないが，ROUGE-2,3 の値は高くなっていることがわかる．これは GCNN によって入力単語のまとまりの特徴量がとれたためと考えられる．bigram や trigram が一般に使われることを考えると妥当な結果といえる．

また stack を積むと精度が悪くなることが見て取れる．これは stack を積むことでネットワークが複雑化し学習がうまくいかないことが原因と考えられる．

帰無仮説を「」と定義するとよって p 値が 0.05 を下回ったので 95%信頼区間で有意差ありと判定された．

生成型要約について比べると，表 2 より，ABS モデルのほうがマルチ Attention モデルより ROUGE 値が高いことが分かる．また生成される要約文の内容についても，ABS モデルに比べてうまく生成できていなかった．マルチ Attention モデルでは文章は生成できているが，内容や意味的に正しい要約がとれていなかった．全文を入力する場合，冗長性があがるために文生成の精度が落ちたためと考えられる．また全文を入力した際は，訓練データ数が少ないこともあり学習が十分にされなかったことも精度低下の要因と考えられる．また，ニュース記事の要約は最初の文が用いられることが多い．平均マルチ Attention モデルが文間の荷重をかけないモデルであり，すべての文を同等に扱うこのモデルの特性が，精度低下の一因とも考えられる．

今後，入力文を全文ではなく最初の 3 文程度を入力としたり，文間に荷重をかけたりすることで平均マルチ Attention モデルより ROUGE の精度が上がると予想する．

5 まとめと今後の課題

ABS モデルと平均マルチ Attention モデルで文章要約をし，両者のモデルの性能の比較をした．全文を同等に扱うマルチ平均 Attention モデルでは単一

表 3: パラメータを変化させた際の ROUGE 値

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-S	ROUGE-L
ABS モデル	0.3127	0.0918	0.0366	0.0962	0.2584
$kernel\_width = 2$	0.3168	0.1102	0.0468	0.0992	0.2612
$kernel\_width = 3$	0.3042	0.1099	0.0521	0.1013	0.2608
$kernel\_width = 4$	0.2953	0.0951	0.0422	0.0946	0.2561
$kernel\_width = 8$	0.2598	0.0732	0.0268	0.0748	0.2250
$stack\_size = 1$	0.3168	0.1102	0.0468	0.0992	0.2612
$stack\_size = 2$	0.2778	0.0736	0.0274	0.0797	0.2334
$stack\_size = 3$	0.2696	0.0761	0.0347	0.0785	0.2268
$stack\_size = 8$	0.2174	0.0384	0.0114	0.0448	0.1814

表 4: タイトル生成例

入力文	兵庫県警川西署は num 日、同署警務課の男性巡査部長（num）が、月末に退職する予定の署員 num 人から預かっていた警察手帳 num 冊を紛失したと発表した。
正解文	巡査部長が警察手帳を num 冊紛失－－兵庫県警川西署
ABS モデル	num 冊の巡査、情報入り紛失
提案モデル ($kernel\_size = 2$,$stack\_size = 1$)	兵庫県警巡査長、男性 num 人を紛失－－兵庫県警
提案モデル ($kernel\_size = 2$,$stack\_size = 8$)	num 歳の女性、num 人を書類送検－－有償運送

文だけを利用する ABS モデルより精度が落ちることが分かった．しかし要約は複数の文を参照して行われるものであるためにマルチ Attention モデルを用いた要約文生成が望ましい．平均マルチ Attention モデルの精度向上のために，ニュース記事の要約は最初の文が用いられることが多いことを考慮して入力文数の制限や，文間の荷重付与が今後必要になると考えられる．それ以外の今後の課題としては，文間の荷重をかけるディープマルチ Attention モデルの実装，ひとつの記事から複数の要約文を生成する手法の考案などが挙げられる．

現在，複数の要約文の生成法として以下の手法を考えている．

- 記事を単純に 3 分割し，それぞれの分割した複数文章を異なる 3 つのモデルを用いて学習させ，それぞれのモデルから 3 つの要約文を生成する手法
- 抽出型要約 (LexRank 法など) で代表文を 3 文取り出して，その 3 文から生成型要約を利用して要約文を生成する，抽出型要約と生成型要約を組み合わせた手法

参考文献

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[3] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*, 2016.

[4] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.