ARG WI2 No.13, 2018
WI2-2018-18

Graph Convolutionにより構文構造を加味したGANによる文章生成手法の提案

澤崎 夏希[†,a]　遠藤 聡志[††,b]
當間 愛晃[††]　山田 孝治[††]　赤嶺 有平[††]

† 琉球大学理工学研究科情報工学専攻　†† 琉球大学工学部知能情報コース

a) *k178577@ie.u-ryukyu.ac.jp*　b) *endo@ie.u-ryukyu.ac.jp*

概要　現在ディープラーニングの発展により様々な問題が解決されているが，その問題の多くは十分なデータ量が確保されており，少量学習データでの問題解決は依然として課題となっている．データ量が少ない場合の対策として，データを増加させるかさ増し手法が用いられる．特に画像分野においては Generative Adversarial Network:GAN を用いた高精度な画像生成手法が注目されている．自然言語の分野においても，GAN を応用し文章を生成する試みが広く行われているが，十分な精度の文章生成を行うのは難しい．原因の一つとして，自然言語生成に用いられる GAN では，多くの場合構文構造は加味されていないことがあげられる．そこで，本論文ではグラフ構造を畳み込む Graph Convolution を用いて，構文構造を加味した上で文章生成を行う手法を提案する．

キーワード　かさ増し, 自然言語, 不均衡データ,GAN,GCN

1　はじめに

現在のディープラーニング技術発展の背景には大規模なデータセットの存在がある．十分なデータが用意出来ない場合にその学習が難しい事が知られており [1]、特にカテゴリ毎のデータ量に差があるケースが問題視されている．カテゴリ毎のデータ量が不均衡な場合、データ数の多いカテゴリを強く学習し，少量データを学習しにくい過学習が発生する．

過学習の問題を防ぐためにデータセットに対して前処理を行いデータ量の不均衡さを解消する手法が用いられる．大別してデータ量を少量カテゴリに合わせるダウンサンプリングと、少量データを増やすかさ増しの 2 つが用いられる．しかしダウンサンプリングは全体のデータ量を減少させてしまうためデータセットの持つ情報を十分に活用出来ない場合がある．かさ増しではデータ量を確保できるため十分な学習が期待できるが，生成したデータに学習データとして特徴を獲得させることは容易ではなく様々な提案がされている．画像処理の分野では、ノイズの付与、輝度勾配の変更、ガンマ値補正等の画像を特徴づける性質を加味した上でかさ増しが行われる．また，ルールベース的な画像処理技術の他に、現在では Generative Adversarial Network:GAN を用いてより抽象的な特徴を含む画像の生成が注目されている．一方で自然言語においては有効なかさ増し手法は確立されておらず、ドメイン毎に人手で特徴を分析し文章生成を行う事が多く負担となっている．そこで自然言語を対象にした機械的なかさ増し手法が必要になる．

GAN よる文章生成手法は画像分野に比べ不安定であり，その原因として学習データの特徴を捉えるのが難しい事があげられる．従来の手法は単語の周辺情報や並びを重視した生成が行われている．本論文では構文構造に注目し，GAN に Graph Convoliutation を用いる事で構文構造を加味し、より学習データの特徴を獲得する文章生成を行う手法を提案する．

2　先行研究

2.1　かさ増し手法

著者らの過去の研究にルールベースでのかさ増し手法 [2] がある．word2vec を用いた類似単語入れ替え，wordnet を用いた類似単語入れ替え，係り受け解析を用いた並列文節入れ替え手法の 3 つを提案した．論文ではニュース記事の不均衡データに対してかさ増し手法を用い，カテゴリ分類実験を行った．その結果カテゴリ情報のかさ増しには成功したが，ドメイン毎にかさ増しのためのルール設計が必要になる．そこで，文章の特徴を機械的に獲得し生成する仕組みが必要になる．

2.2　GAN

GAN は Goodfellow らが提案した判別器 D と生成器 G を協調学習させることで生成を行うモデル [3] である．D は入力が学習データか（真判定）生成データか（偽判定）を判別し，その判別結果を元に G の学習を行い学習データに近いデータを生成する．学習が進んだ G は D に判別されにくいデータを生成し，そのデータと学習データからさらに D が学習を行い判別精度を高める．これにより機械的に特徴を獲得し生成を行う事ができる．

一方で学習が不安定という問題があり，G は真判定さ

れた生成データに近く，偽判定された生成データから遠いデータを生成するように学習する．このため D の出力が真偽どちらかに大きく偏ってしまうと G の更新のための勾配が得られず G の学習が止まってしまう．その結果同一のデータが大量に出力されてしまう mode-collapse 問題が課題となっている．

2.3 Sequence Generative Adversarial Nets

Sequence Generative Adversarial Nets:SeqGAN は Yu らが提案した手法で自然言語に対し GAN を適用したモデルである [4]．判別器 D に CNN を用い，生成器 G に LSTM を用いている．さらにモンテカルロ探索を用い現在の文章の状態と次の単語の選択に対する報酬を加えることで，強化学習の枠組みを利用した文章生成を行う．G では LSTM を用いて次の単語を予測し，その予測した結果から D が報酬を出力する．判別結果の報酬を G に渡すことで学習を行い，D に判別されにくい文章の生成を行う．この生成器 G，判別器 D，報酬関数という構成は自然言語の生成によく用いられ，本論文でも同様の構成を用いる．

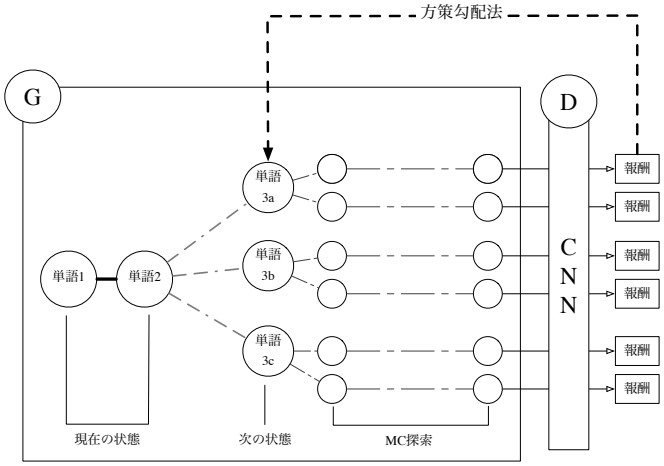

図 1　SeqGAN

2.4 TextGAN

Yizhe らの提案した GAN に mode-collapse 問題への対策を行った文章生成モデルである [5]．mode-collapse 問題への対策として学習データと生成データの分布の距離計算に Maximum Mean Discrepancy:MMD[6] を用いている．これにより勾配の消失が起こりにくくなり，生成器 G が学習データの分布を獲得しやすくなる．TextGAN と比較することで学習データの分布を獲得出来ているかを測ることが出来る．

2.5 Texygen

Yaoming らは文章生成に用いられる GAN を比較する実験 [7] を行った．実験に用いた GAN は，SeqGAN, GSGAN[8], MaliGAN[9], RankGAN[10], LeakGAN[11], TextGAN の 6 つである．比較は学習データと生成データとの平均 BLEU スコア [12]，及び生成データ内で平均 BLEU スコアで行っている．学習データと生成データの BLEU スコアが大きい場合，学習データの特徴を獲得出来ていると評価され，生成データ内での平均 BLEU スコアが小さい場合多様性のある文章生成が行えていると評価した．一方で BLEU スコアのみでは精度指標として不十分であることも述べている．本論文では BLEU スコアを用いた評価方法に cos 類似度による比較を加え，かさ増し手法の評価を行う．

2.6 GraphVAE

GraphVAE は Martin らが提案したグラフを生成するためのモデル [13] である．Martin らは Variational Autoencoder:VAE を用いてグラフ構造を確率的に扱うことで学習しやすくしている．VAE は学習データの分布を仮定することで生成モデルの学習を安定させるもので，生成モデルの持つ mode-collapse 問題を軽減する事が出来る．

2.7 MolGAN

MolGAN は Nicola らが提案したモデル [14] であり，GAN の判別器 D に Relational Gprah Convolution Netwark:R-GCN[15] を採用している．生成器 G には GraphVAE を用い，生成したグラフ構造が正しければ報酬を与えることでより安定した生成を行う工夫がされている．小規模なグラフの生成に有効であるとしており，大規模なグラフの生成は難しいことを示している．

3 要素技術

3.1 Relational Graph Convolution Nerwork

R-GCN はグラフ内のノードの近傍情報を畳み込むモデルである．ここでいう近傍情報とは，あるノードに注目したときの注目ノードに接続されているエッジの種類，接続されたエッジの先にあるノードの持つ特徴，注目ノード自身の持つ特徴である．この近傍情報を内包した抽象的な特徴を獲得する仕組を畳み込みと呼ぶ．第 1 層の R-GCN 層で近傍 1 の情報を，第 2 層で近傍 2 の情報を畳み込み，層の深さに比例して大域的な情報を獲得することが出来る．図 2 では，グラフ A が A' に畳み込まれている様子を表している．

3.1.1 R-GCN の出力例と近傍情報

R-GCN に入力される近傍情報としてグラフ構造から隣接行列と特徴行列を用いる．隣接行列はエッジの接続の有無を表し，特徴行列はノードの持つ特徴を行ベクトルで表している．隣接行列はエッジの種類毎に用意され，図ではグラフ A が持つエッジ $A\alpha$ とエッジ $A\beta$ が隣接行列 A として表されていることを示す．特徴行列は

ノードの持つ特徴であり，原子や単語の種類，個人の情報など様々な特徴を表現することが可能である．図では 1 つのノードにつき 2 つの特徴を持つ場合の特徴行列 F を示している．

R-GCN はノード単位で行われる．まずノード a に対して R-GCN が用いられ，近傍情報が畳み込まれた a' が得られる．この a' の持つ特徴に近傍情報が畳み込まれている．次に b に対して R-GCN が用いられ b' が得られ，同様に c' と d' が得られる．こうして得られたノード a'〜d' を元に特徴行列 F' が得られる，畳み込まれたグラフ A' が得られる．A' は中間層となり，入力となるグラフと同じ大きさである．

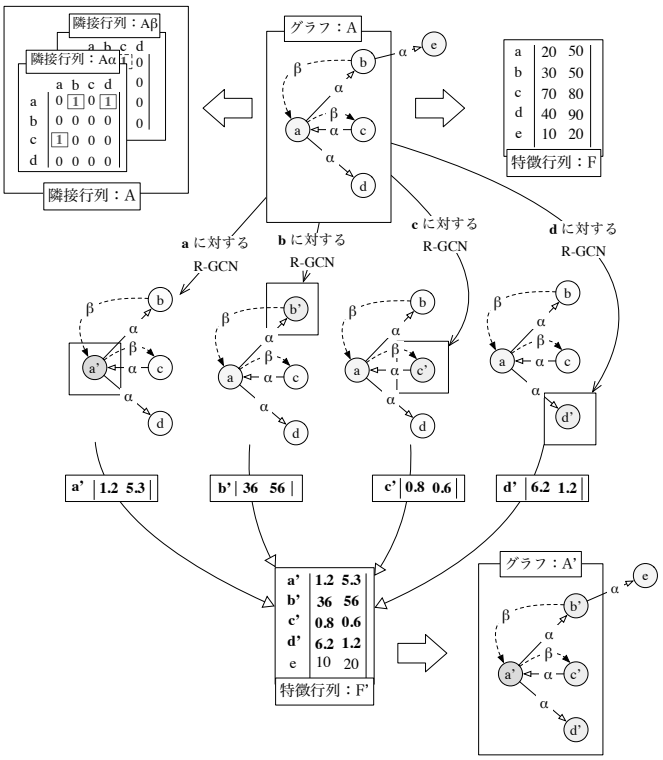

図 2 R-GCN の出力例と隣接行列，特徴行列

3.1.2 R-GCN の詳細な処理

次に R-GCN の詳細な処理を図 3 に示す．ノード a に注目した時，近傍情報は左の矩形内のノードとエッジになる．この時，エッジ毎にグラフを部分グラフに分けて計算を行う．二種類のエッジ α，β がある時，部分グラフであるグラフ $A\alpha$，グラフ $A\beta$ が得られる．また自身のノード情報も加味するため，自己ループのグラフ $Aself$ が得られる．さらにそれぞれのグラフは，a に対する接続方向も加味する．エッジ α の場合，入力が 1 つということを表すノード集合 α_In が得られ，ノード a からの出力が 2 つであるということを表すノード集合 α_Out，a が得られる．この得られたノード集合にそれぞれ第 I 階層目の重み行列 WI をかけ合わせ，合計したものを活性化関数に入れることで近傍情報を畳み込んだ出力ノード a' が得られる．この時の自己ループは固有の重み $W0$ で計算することで注目ノードの特徴の消失を防ぐ．以上の処理を全てのノードについて行うことでグラフ構造の畳み込みを行う．

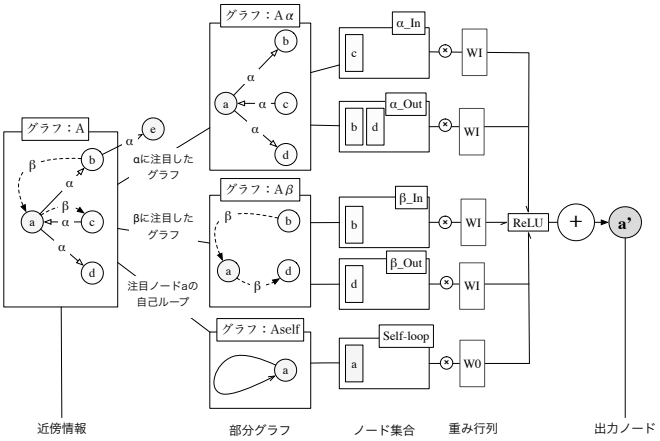

図 3 R-GCN の詳細な処理

4 提案手法

本研究では，MolGAN をベースに構文情報を加味した新たな GAN を提案する．今回は構文情報として基本的な文構造である係り受け情報を採用する．単語ベースでは係り受け情報の複雑化と語彙の増加により，グラフが大規模になり生成が難しくなることが考えられる．そこで係り受け解析を分節ベースで行うことでグラフの規模の縮小を図った．ノードの特徴については，分節毎に設定した分節 ID を用いる．これは単語ベースにおける語彙に相当する情報になる．

判別器 D には R-GCN を用い，生成器 G には RNN と GraphVAE を用いる．判別器 D に R-GCN を使うことで，係り受け情報といった構文情報を加味できるのではないかと考えた．報酬には生成データ間の平均 BLEU スコアが低い場合に生成文章のバリエーションが多いと考え報酬を与える．

4.1 生成器 (Generater)

今回用いる生成器 G には，GraphVAE と RNN を用いる．GraphVAE を用いることによりグラフ構造を確率的に扱う事が出来る．分布を仮定して生成を行うため生成データの多様性を維持できる特徴がある．

RNN を用いた生成器 G では系列情報を元に文章を生成する事ができ，文章の生成において広く使われる生成器である．G は D からの判断結果と報酬を受けて学習を行い，グラフの構造と生成のバリエーションを加味した生成を行う．G の学習が D と比較し進みすぎると D が偽判定を行えず，同一のデータが多く生成されてしまう

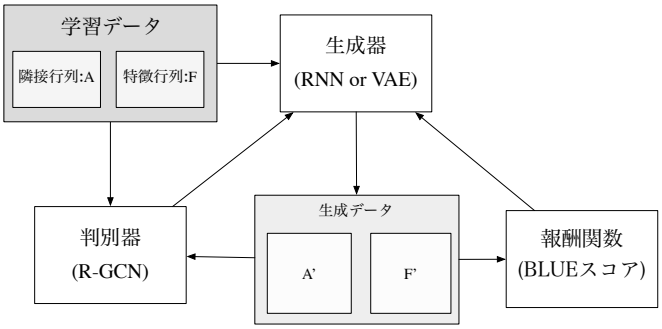

図 4　提案手法

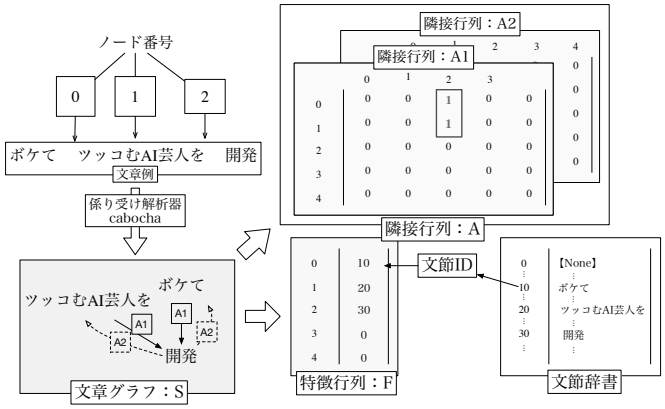

図 5 文章生成におけるノードとエッジの設定

可能性があるため学習回数の比率の調整が必要になる．

4.2　判別器 (Discriminator)

隣接行列と特徴行列を入力にした R-GCN を判別器 D に用いる．これによりグラフの構造からみた真偽判定を行う事ができ，生成器 G は構文情報を加味した文章生成が可能になる．そのため隣接行列と特徴行列が上手くグラフの構造を表していることが重要になる．

4.3　ノードとエッジの設定例

文章に係り受け解析を行い，分節毎の係り受け情報を特徴としたグラフを作成する (図 5)．図では文章に係り受け解析を行い，係り受け関係から隣接行列と特徴行列得るまでの工程を表している．文章は cabocha[16] を用いて文節単位に分割され，その文節毎に係り受け解析行う．解析によって得られた係り受け情報から文章グラフ S から隣接行列と特徴行列を得る．係り受けの方向で別のエッジとして獲得される．図ではエッジ $A1$ を係り受けする関係，エッジ $A2$ を係り受けされる関係である．ここから隣接行列 $A1$ と $A2$ が得られる．

ノードの特徴には文節の ID を持たせている．これは学習データから得られた文節辞書を使用する．ここで 0 は空ノードを表す【None】が記録されている．この文節 ID がノードの特徴となり，この文節 ID を元に生成されたグラフを文章に変換する．ノード数は文節の数によって決定され，実験では学習データから最大の文節数であった 9 をノード数として採用している．

5　実験

本論文では，実験として類似度の比較（図 6）を行う．学習データと生成データの類似性，生成データの多様性を比較する．ランダムに 1000 件抽出した学習データと生成データの BLEU スコアと cos 類似度の平均を用いることで学習データとの類似性を計測することが出来，生成データ同士の BLEU スコアと cos 類似度を用いて生成データの多様性を計測出来る．

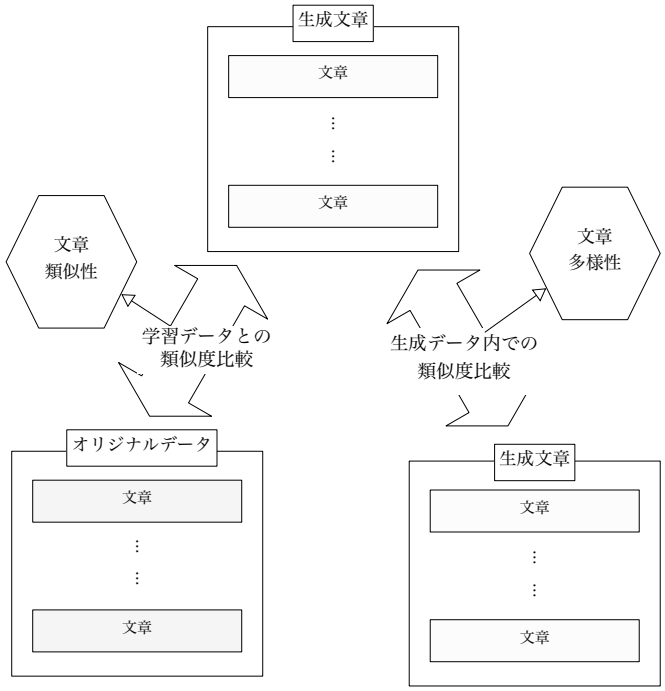

図 6 類似度測定実験と多様性測定実験

5.1　データセット作成

Yahoo!ニュース [17] のタイトルをカテゴリ毎に獲得しデータセットとした．8 つあるカテゴリは最大 55, 0180 件，最小 14, 813 件とデータ数に大きな差がある不均衡データであり分類が難しく，短文であるため文章生成を行いやすい特徴がある．生成はルールベースによるかさ増し，SeqGAN，TextGAN，生成器 G にに GraphVAE を用いた提案手法．G に RNN を用いた提案手法の 5 つの手法を用いて行う．

IT カテゴリのデータ 14, 880 件を学習した際の文章生成の例を図（7）に示す．ルールベースによるかさ増し例では対象となった単語以外が生成前の文章であるため，変化が少ない生成になる．文章としては成立したものを生成しやすいが，生成文章のバリエーションが少

【学習データ】

・部下に Twitter やめろ ありか

・暴走系ゆるキャラ 動画が話題

【ルールベースによるかさ増し例】

・知りが抑えたい費用はこれの携帯

・グリー「サン値引き中小 2 位はキャノン

【Seqgan によるかさ増し例】

・巨人も障害「ラブ残しに終了

・不正攻撃、q6 に「ソフトソフト

【Textgan によるかさ増し例】

・ヒ素東部層地に

・氏病のの「「「「「「「「「「

【提案手法（GraphVAE）によるかさ増し例】

・外来ホヤ、公開拡大有効?治療法はメガソーラー構想清水建すべて過剰摂取した小保方氏悲しい栄養

・銀河周辺の排せつ物発電の飛行準備ベッドで歯止め回復へ賢さをふん化石 82 日ぶり捕獲ワクチン健康被害母親の

【提案手法（RNN）によるかさ増し例】

・ネタ満載 AI 採用どうツクモネットと Google メモ作成アプリ開発かもうからず?展開任天堂スマホ攻略苦戦日本好き新記録

・不正送金情報ウクライナにスマホ新指針総務省の表情記念に Web 文字おこしは線引き物議法的にはデータの導入を e スポーツに

図 7 学習データと文章生成結果

ない．SeqGAN によるかさ増しでは変化が大きくバリエーションのある生成が可能であるが，単語間の関係はあまり獲得出来ておらず，同じ単語が繰り返されている．TextGAN は SeqGAN よりも顕著に単語の繰り返しが見られ，学習データの特徴を十分に獲得出来ない．提案手法による文章生成はどちらも複数の文章が含まれているようなものが生成されている，これは文節間の係り受けを獲得出来てると見ることが出来るが，文章全体の構造は十分に獲得出来ていない．

5.1.1 ルールベースによるかさ増し

今回用いるルールベースによるかさ増しは，単語間の類似度を用いて行う．入れ替える対象となった単語（単語 C）の単語ベクトルを用い cos 類似度により，上位 10 件の類似単語からランダムに選択された単語を入れ替えることでかさ増しを行う．単語ベクトルは日本語版 Wikipedia で学習済みのモデルを用いて獲得する．未知語は入れ替え対象とならない．

5.2 実験結果

実験では学習データと生成データの類似性と，生成データの多様性の 2 つを評価する．学習データと生成データの類似度を測ることで，どれだけ学習データの特徴を加味した文章が生成できたかを，生成データ同士の

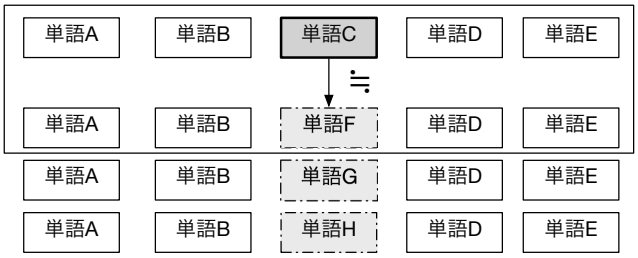

図 8 ルールベースによるかさ増し

類似度を測ることで生成された文章がバリエーションのある文章かを評価する．類似度計算は BLEU スコアと cos 類似度を用いる．学習には IT カテゴリ 14,880 件，科学カテゴリ 14,813 件をそれぞれ用いた．

5.2.1 学習データとの類似度比較実験

表 1 は学習データと生成データの類似度を計算したものである．数値が高いほど類似性が高いと見る事ができ，元の文章に近い文章が生成できているといえる．cos 類似度に用いる文書ベクトルは学習データにかさ増しデータを加えたもので学習を行った．類似度が高いものを太字で表している．cos は cos 類似度を表す．

表 1 学習データとの類似度比較結果

	IT		科学	
	BLEU	cos	BLEU	cos
ルールベース	0.0373	0.3277	0.0516	0.2950
SeqGAN	0.0405	0.1857	0.0548	0.2939
TextGAN	0.0359	0.2781	0.0516	0.2282
提案手法 (GraphVAE)	0.0971	**0.3460**	**0.1127**	0.3313
提案手法 (RNN))	**0.0975**	0.3286	0.1123	**0.3328**

IT カテゴリの BLEU スコアと科学カテゴリの cos 類似度に対して G に RNN を用いた提案手法が，IT カテゴリの cos 類似度と科学カテゴリの BLEU スコアに対しては G に GraphVAE を用を用いた提案手法が最も学習データとの類似度が高かった．これは提案手法が比較的元データの特徴を獲得できたといえ，その特徴として係り受け構造が有用であることを示している．また提案手法は IT カテゴリと科学カテゴリの両方で BLEU スコアが他の 2 つの手法と比べて高い．これは単語単位で見た類似性が高いということであり，学習データで係り受け関係にあった文節を獲得出来たといえる．

5.2.2 生成データ内の類似度比較実験

表 5.2.2 は生成データ内の類似度を計算したものである．ベースラインとして学習データの類似度を算出した．これにより現実的なデータセットの類似性との比較を行うことが出来る．類似度は数値が低いほど多様性がある，すなわちバリエーションのある文章が生成できているといえる．データ毎の多様性を見るため，文書ベク

トルはデータ毎に学習している．類似度の低いものを太字で表している． 学習データの類似度はカテゴリごと

表 2 生成データ内の類似度比較結果

	IT		科学	
	BLEU	cos	BLEU	cos
学習データ	0.0357	0.0711	0.0516	0.0193
ルールベース	0.1826	**0.4396**	0.1886	0.5910
SeqGAN	0.0579	0.4456	**0.0598**	0.5882
TextGAN	**0.0415**	0.4439	0.1742	**0.5846**
提案手法 (GraphVAE)	0.1893	0.4486	0.1897	0.5912
提案手法 (RNN)	0.1898	0.4593	0.1826	0.5975

に差がある．特に BLEU スコアが科学カテゴリより IT カテゴリの方が低い．これは使用される単語の語彙が広いことを表し，企業名や製品名など短い期間だけ使用される単語があると予想される．一方で科学分野では比較的長期間使用される専門用語が多いことが予想される．

次に生成文章の類似度を見る．IT カテゴリの BLEU スコアでは SeqGAN と TextGAN が優位を示した．これは SeqGAN が探索木を用いるため，バリエーションのある文章生成が可能なことが要因だと考えられる．また TextGAN は mode-collapse 問題への対策が行われておりその効果が表れていると言える．一方で提案手法の多様性は低く，これは係り受け関係にある文節のバリエーションが少ないことが原因として考えられる．cos 類似度を見るとどの手法も大きな差は無く，学習データと比較すると多様性に大きな課題がある．

6 まとめ

本論文では構文構造を加味した文章生成を行うために GAN に Graph Conbolution を採用したモデルを提案した．入力に用いる隣接行列は係り受けの方向，特徴行列は分節 ID を用い，生成結果を分析するために類似度比較実験を行った．類似度実験では他の生成手法よりデータの特徴を加味した生成が行えていることが示されたが，同時に多様性の消失も見られた．現在の文章生成手法を用いて現実世界のデータと同程度のバリエーションを持たせるのは難しく，かさ増しとして扱うには課題がまだ多い．

改善案としてはまず報酬関数の設定やエッジやノート特徴の追加が考えられる．また生成するグラフは学習データの語彙に大きく依存するため，文章ではなくより抽象的な文構造の生成などのアプローチが考えられる．その際は wordnet などの人手でのコーパスも併用し，機械的な生成手法とルールベース的な生成手法を複合した手法を提案する必要がある．

参考文献

[1] He, Haibo, and Edwardo A. Garcia. ”Learning from imbalanced data.” IEEE Transactions on knowledge and data engineering 21.9 (2009): 1263-1284.

[2] 澤崎 夏希, 遠藤 聡志, 當間 愛晃, 山田 孝治, 赤嶺 有平. ”量的不均衡データに対する学習精度改善のための文書かさ増し手法” 第 11 回 Web インテリジェンスとインタラクション研究会 (2017)

[3] Goodfellow, Ian, et al. ”Generative adversarial nets.” Advances in neural information processing systems. 2014.

[4] Yu, Lantao, et al. ”SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.” AAAI. 2017.

[5] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. Adversarial Feature Matching for Text Generation. arXiv preprint arXiv:1706.03850 (2017).

[6] Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. JMLR, 2012.

[7] Zhu, Yaoming, et al. ”Texygen: A Benchmarking Platform for Text Generation Models.” arXiv preprint arXiv:1802.01886 (2018).

[8] Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. arXiv preprint arXiv:1611.04051 (2016).

[9] Tong Che, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-Likelihood Augmented Discrete Generative Adversarial Networks. arXiv preprint arXiv:1702.07983 (2017).

[10] Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial Ranking for Language Generation. arXiv preprint arXiv:1705.11001 (2017).

[11] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Long Text Generation via Adversarial Training with Leaked Information. arXiv preprint arXiv:1709.08624.

[12] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002) BLEU: a method for Automatic Evaluation of Machine Translation. ACL.

[13] Simonovsky, Martin, and Nikos Komodakis. ”GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders.” arXiv preprint arXiv:1802.03480 (2018).

[14] De Cao, Nicola, and Thomas Kipf. ”MolGAN: An implicit generative model for small molecular graphs.” arXiv preprint arXiv:1805.11973 (2018).

[15] Schlichtkrull, Michael, Kipf, Thomas N, Bloem, Peter, Berg, Rianne van den, Titov, Ivan, and Welling, Max. Modeling relational data with graph convolutional networks. arXiv preprint arXiv:1703.06103, 2017.

[16] CaboCha/南 瓜: Yet Another Japanese Dependency Structure Analyzer. ”https://taku910.github.io/cabocha/”

[17] Yahoo!ニュース ”https://news.yahoo.co.jp”