原著論文

MolGANの拡張による文章グラフを用いた文章生成手法の提案†

澤崎　夏希*1・遠藤　聡志*2・當間　愛晃*2・山田　孝治*2・赤嶺　有平*2

深層学習によって様々な分類問題が解決されているが，分類カテゴリ毎のデータ量が不均衡な問題を扱う場合，多くの課題がある．不均衡データへの対策として，少量カテゴリのデータ量を増加させ均衡化する手法がある．これをかさ増しと呼び画像処理分野ではノイズの付与や回転による方法が一般的である．最近では Generative Adversarial Network: GAN による画像生成手法を用いる場合がある．一方で，自然言語処理の分野では有効なかさ増し手法はいまだ確立されておらず，人手によるかさ増しが行われている．人手によるかさ増しではルールの設計など負担が大きく，機械的なかさ増し手法が必要となる．しかし，文章生成における機械的なかさ増しは画像生成に比べ不安定である．これは文章の特徴獲得の難しさが原因だと考えられる．そこで本論文ではグラフ情報に注目した機械学習による文章生成手法を提案する．CaboCha によって生成されたグラフ情報を Graph Convolution により畳み込み処理する．提案する GAN により生成されたかさ増し文章を 3 つの計算実験により評価し有効性を示した．

キーワード：自然言語のかさ増し，不均衡データ，GAN

1. はじめに

現在深層学習によって扱われている問題の多くは，分類問題を解く形で解決されている．一方で分類問題を解く際には，各カテゴリにある程度均衡なデータ量が用意されていることが前提とされており，カテゴリ毎のデータ量が不均衡な場合，学習が難しい事が知られている [1]．不均衡データを入力として学習した場合，データ量の多いカテゴリを強く学習し，データ量の少ないカテゴリは学習しにくくなる．これは過学習の一種として知られている．

こ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サンプリング手法が用いられる．大別してデータ量を少量カテゴリに合わせて削減するダウンサンプリング手法と，少量カテゴリのデータ量を増加させるかさ増し手法の 2 つが存在する．ダウンサンプリングは全体のデータ量を削減し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．一方，かさ増し手法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，学習データの特徴を獲得した生成を行う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，画像の回転，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 [2] を用いた，より抽象的な特徴を含む画像のかさ増しが行われている [3].

一方で自然言語処理においては有効なかさ増し手法は確立されておらず，人手で作成されたルールを用いて行われる [4]. レビューや SNS などのデータセット毎に異なるルールが必要な場合があり，人手で作成することは大きな負担になる．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．機械的なかさ増しを行う上で必要なのは生成データの語彙空間であり，テキストに対して GAN を用いることで多様な文章生成が行えると考えた．しかし GAN による文章生成手法は画像分野に比べ不安定である．その原因として自然言語では学習データの特徴を捉えるのが難しい事があげられる．従来の手法で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 [5] が，これらの特徴だけでは文章としての特徴を十分に獲得できない．

本論文では，CaboCha [6] を用いて生成したグラフ情報を付与する事を提案する．グラフの特徴は Graph Convolution によって畳み込まれる．提案する GAN により生成されたかさ増しした文章を生成実験，類似度比較実験，分類実験により有効性を評価する．

2. 先行研究

2.1 人手による自然言語のかさ増し手法

著者らの過去の研究にルールベースでのかさ増し手法 [7] がある．word2vec [8] を用いた類似単語入れ替え，wordnet [9] を用いた類義語入れ替え，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．この研究では不均衡データであるニュース記事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．データ量を 200 件かさ増しした結果分類精度の改善が見られたが，630 件のかさ増しを行った結果分類精度が低下した．これは類似単語や文節の入れ替えなどの，人手で設定

† Sentence Generation Method by Extension of MolGAN Using Sentence Graph
Natsuki SAWASAKI, Satoshi ENDO, Naruaki TOMA, Koji YAMADA, and Yuhei AKAMINE

*1 琉球大学工学部理工学研究科情報工学専攻
Graduate School of Information Engineering, University of The Ryukyus
*2 琉球大学工学部工学科知能情報コース
Information Technology Intelligent System, University of The Ryukyus

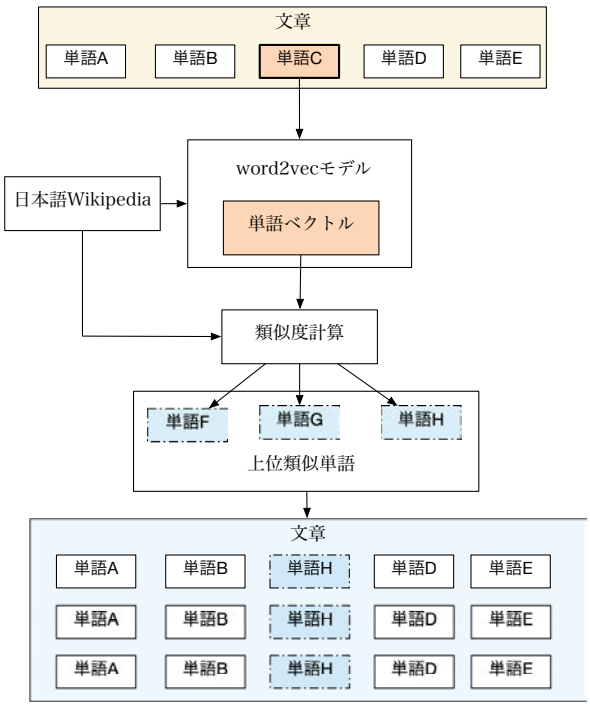

図 1　類似単語入れ替えによるかさ増し手法

した特徴量では十分に多様性のあるデータの生成が行えないことが原因である．多様性を確保するためににはドメインを分析し，再び人手で特徴を設定する必要がある．本論文では提案手法との比較のために，word2vec による単語入れ替え手法をルールベースのかさ増しとして用いている．

2.1.1　word2vec による単語入れ替え手法

ルールベースのかさ増し手法として，単語間の類似度を用いた手法がある（図 1）．これは入れ替える対象となった単語の単語ベクトルから cos 類似度を計算し，上位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手法である．特徴として，周辺単語の類似度を特徴としたかさ増しが可能だが，文章の大部分は変化しないことが挙げられる．単語ベクトルは日本語版 Wikipedia で学習した Word2vec 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．

2.2　GAN

GAN は Goodfellow らが提案した判別器（Discriminator）と生成器（Generator）を敵対的に学習させることで画像の生成を行うモデルである [2]．判別器は，入力が学習データであるか生成データであるかを判別する．生成器は判別結果を元に学習を行い学習データに近いデータを生成する．学習が進んだ生成器は判別器に判別されにくいデータを生成し，さらに判別器が生成データと学習データの双方から学習を行い判別精度を高める．このような敵対的学習により抽象的に特徴を獲得し生成を行う．結果として生成器は学習データと判断された生成データに近いデータを生成するように学習する．

しかし GAN による学習は不安定であり，生成データを入力とした際に判別器の出力がどちらかに大きく偏ってしまうと，重み更新のための勾配が得られず生成器の学習が行われない．その結果，同一のデータが大量に出力されてしまう mode collapse 問題が生じ，これはかさ増しにおいて大きな問題となる．グラフ構造を学習させることは適切な特徴を獲得することにつながり，勾配消失を防ぐ mode collapse 問題への対策となりうる．

2.2.1　SeqGAN

Sequence Generative Adversarial Nets: SeqGAN は Yu らが提案した手法で，自然言語に対し GAN を適用した文章生成モデルである [10]．判別器に CNN を用い，生成器に 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，次の単語の選択結果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．生成器では LSTM を用いて次の単語を予測し，その予測結果を元に判別器が報酬を出力する．判別結果の報酬を生成器に渡すことで学習を行い，判別器に判別されにくい文章の生成を行う．この生成器，判別器，報酬関数という構成は自然言語の生成によく用いられ，本論文でもこれら 3 つを用いた GAN を設計する．

2.2.2　Texygen

Yaoming らは文章生成に用いられる複数の GAN を比較する実験 [11] を行った．実験に用いた GAN は，SeqGAN，GSGAN [12]，MaliGAN [13]，RankGAN [14]，LeakGAN [15]，TextGAN [16] の 6 つである．提案手法を評価するため，Texygen に含まれる SeqGAN と TextGAN を比較対象として用いている．SeqGAN と比較することで従来の文章生成手法と提案手法の多様性を評価し，TextGAN との比較を行うことで mode collapse 対策後の文章生成手法と提案手法の多様性を評価する事が出来る．評価では単語単位の類似性を見るために，平均 BLEU スコア [17] を用いる．学習データと生成データの類似度が大きい場合，学習データの特徴を獲得出来ていると評価し，生成データ内での類似度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される．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加え，文脈としての類似度を見るため，文書ベクトルの cos 類似度による比較を行う．

2.3　グラフ畳み込み手法

2.3.1　R-GCN

Relational Graph Convolution Nerwork: R-GCN はグラフからノードの近傍情報に注目した特徴を獲得するモデルである [18]．ここで近傍情報とは，あるノードに注目したときの周辺のノード，ノードに接続されたエッジ，注目ノード自身の持つ特徴の 3 種類である．この近傍情報を内包した抽象的な特徴を獲得する仕組みをグラフ畳み込みと呼ぶ．

図 2 にグラフが畳み込まれる様子を表す．グラフ畳み込みは入力されたグラフから注目ノード a の近傍情報についての部分グラフ A を作成し，そこからさらにエッジの種類ごとに部分グラフ $(A_\alpha, A_\beta, A_{self})$ を得る．このエッジ毎のノード集合と畳み込みの為の重み $W1$ を計算し，最終的に近傍情報を畳み込

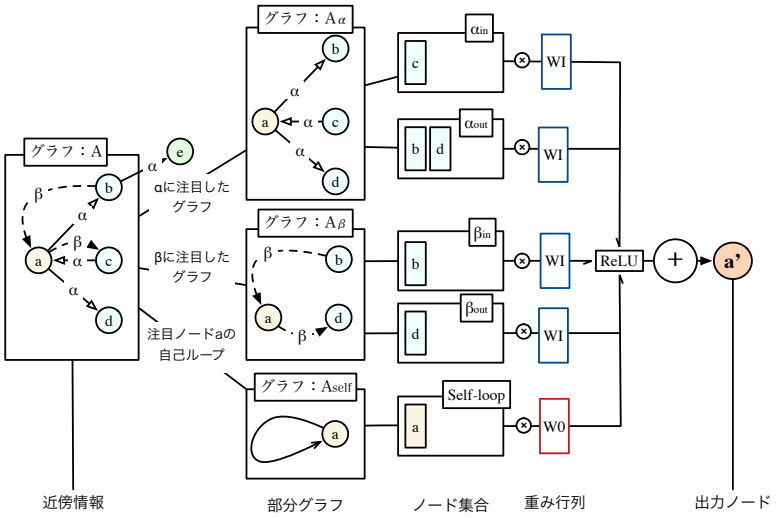

図２　R-GCN によるグラフの畳み込み

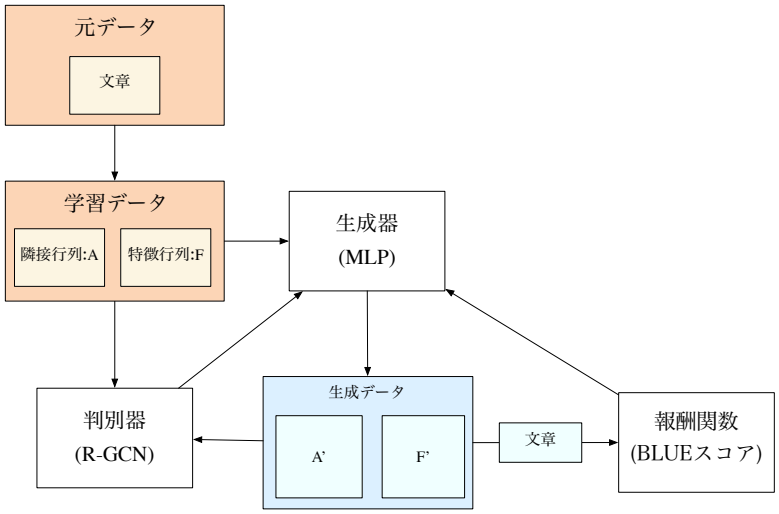

図３　提案手法

まれた特徴ノード a' が得られる．また，注目ノード自身の特徴は別の重み $W0$ により計算される．この処理を全てのノードに対して行うことで，近傍情報が畳み込まれたグラフを得る事が出来る．

2.3.2　MolGAN

MolGAN は Nicola らが提案したグラフを生成するモデル [19] であり，GAN の判別器に R-GCN を採用している．生成器には 3 層の多層パーセプトロンを用い，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．Nicola らは大規模なグラフの生成は難しいとし，小規模なグラフである構造式の生成問題を扱っている．提案手法におけるグラフ構造の扱いは基本的に MolGAN に従う．

3. 提案手法

本論文では，グラフ情報を加味した GAN による文章生成手法を提案する（図 3）．判別器には R-GCN を用い生成器に多層パーセプトロン（Multilayer perceptron: MLP）を用いた GAN で，文章構造はグラフとして表現することが出来るため，グラフ構造からみた判別を R-GCN を用いて行う事で文章情報を加味した生成を行う．CaboCha によって獲得されたグラフ構造を用いた文章生成を行う．この時，グラフの規模縮小のために文節単位での生成が行われる．グラフはエッジの種類と接続を表す隣接行列とノードの特徴を表す特徴行列で表される．

3.1　データセット

データセットには Yahoo!ニュース [21] から，カテゴリごとにスクレイピングしたニュースタイトルを用いる．スクレイピング期間は 2008/10～2018/11 であり，オリジナルのデータセットを表 1 の不均衡データに示す．8 つあるカテゴリは最大 55, 018 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，短文であるため文章生成を行いやすい特徴がある．

表 1 実験に用いるデータセット

カテゴリ名	不均衡データ	均衡データ	かさ増し 不均衡データ	かさ増し 均衡データ
国際	30,388	14,000	30,388	30,000
国内	45,410	14,000	45,410	30,000
経済	30,842	14,000	30,842	30,000
エンタメ	37,468	14,000	37,468	30,000
スポーツ	55,018	14,000	55,018	30,000
IT	14,880	14,000	**30,000**	**30,000**
科学	14,813	14,000	**30,000**	**30,000**
地域	46,982	14,000	46,982	30,000
合計	275,801	112,000	306,108	240,000

表 2 CaboCha の係り受け精度

カテゴリ名	文節数 3 以上	係り受け解析 成功 (%)	係り受け解析 失敗 (%)	文節区切り 失敗 (%)
国際	12,142	52	13	35
国内	20,628	45	21	34
経済	11,547	43	23	34
エンタメ	15,297	55	15	30
スポーツ	23,826	44	18	28
IT	14,083	56	15	29
科学	18,061	46	16	38
地域	21,265	56	8	36

3.2 CaboCha の精度測定

本研究ではグラフ構造の獲得に係り受け解析器である CaboCha [6] を用いている．CaboCha の係り受け精度はデータにより差がでるため，yahoo ニュースデータセットに対する係り受け精度の確認を行った．表 2 は，文節を 3 つ以上もつ文章からカテゴリ毎にランダムに 100 文選択し，係り受け精度を確かめた結果である．表 2 より今回用いるデータに対して CaboCha は一定の係り受け精度を持つもことが分かったものの，係り受け構造の有効性を議論するには十分な精度でないと考えられる．一方で，CaboCha が生成するグラフ構造は文章毎に様々であり，その特徴を表したものだと考えられる．本研究では，CaboCha をグラフ構造生成器と位置づけ文章のグラフ化を行うこととする．また，CaboCha での文節区切りも失敗することが示されたが，本研究では CaboCha で獲得した意味の単位として文節という表現を用いることとする．

3.3 CaboCha による文章のグラフ化

図 4 では文章に CaboCha を用い獲得した文章グラフから隣接行列と特徴行列を得るまでの工程を表している．文章は CaboCha 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を行う．解析によって得られた文章グラフ S から，隣接行列と特徴行列を得る．また，グラフは方向で区別され，それぞれ別のエッジとして獲得される．図ではエッジ A_1 を接続する関係，エッジ A_2 を接続される関係とし，隣接行列 A_1 と A_2 が得られる．ここで隣接行列の大きさは，学習されるグラフの最大ノード数から決定される．今回は学習データ内の最大文節数であった 9 を最大ノード数として採用している．ノードの特徴として文節の ID を設定した．文節 ID は学習データから得られた文節辞書から獲得され，単語の語彙に相当する特徴になる．

3.4 生成器（Generater）

生成器は 3 層の多層パーセプトロンを用いて隣接行列と特徴行列を生成する．グラフを生成する際に用いられる Graph-VAE [20] を元にした手法で，学習データの潜在変数を正規分布と仮定することで安定した生成を行うことが出来る．生成したグラフ構造は判別器によって評価され，その勾配を元に生成器が学習を行う．結果として，生成器は学習データに近いグラフ構造を生成するようになる．

3.5 報酬関数と判別器（Discriminator）

提案手法では，多様性が高いほど報酬が高くなる報酬関数と，グラフ構造を畳み込む R-GCN を判別器として用いている．多様性の高さは生成データ間の類似度で評価され，類似度が低いほど多様性が高いと評価される．判別器はグラフ構造を表現した隣接行列，特徴行列を入力とし，R-GCN を用いる事でグラフ情報を含む文章ベクトルを獲得する．この文章ベクトルを元に真偽判定を行う事で，グラフ情報を加味した文章生成が可能になる．報酬関数 R_ψ は以下の式 (1) で表される．

$$R_\psi = \frac{\sum_{i=1}^{n} BLEU_{self}(S, s_i)}{n} \tag{1}$$

ただし，S は n 個の文章集合全体を，s_i は i 番目の文章であることを表し，$BLEU_{self}$ は文章集合と文章の BLEU スコアをそれぞれ比較しその平均値を返すことを表す．

4. 実験

提案手法を，生成実験，類似度比較実験，分類実験の 3 つで評価する．生成実験では繰り返し単語を含む文章数から自然な文章らしさと，特徴単語の一致度から学習データとの類似性を他の生成手法との比較により評価する．類似度比較実験では生成したデータの特徴を類似度から評価する．学習データと生成データとの比較において，類似度が高い程学習データの特徴を獲得できたといえる．また生成データ同士の類似度比較では，類似度が低いほど多様性があるといえる．分類実験では提案手法をかさ増しとして用いた際の分類精度の変化を評価する．また不均衡データに対しての分類能力を全体精度だけで評価することが難しいため，指標としてカテゴリ毎の適合率，再現率，F 値，カテゴリ毎の F 値から算出した標準偏差を用いる．

4.1 生成実験

文章生成はルールベースによるかさ増し [7]，単語単位で学習を行った SeqGAN，TextGAN，文節単位で学習を行った SeqGAN，TextGAN，提案手法 6 つの手法を用いて行う．SeqGAN と TextGAN について 2 通りの学習を行っているのは，提案手法では学習を文節単位で行うためグラフ情報の影響をみるため SeqGAN と TextGAN についても文節単位の学習を行った．科学カテゴリのデータ 14,813 件を学習した際の文章生成の例を図 5 に示す．実験に用いたパラメータを表 3 に示した．

ルールベースによるかさ増しでは対象となった単語のみを変化させるため，自然な文章になりやすいが変化が少ない．GAN

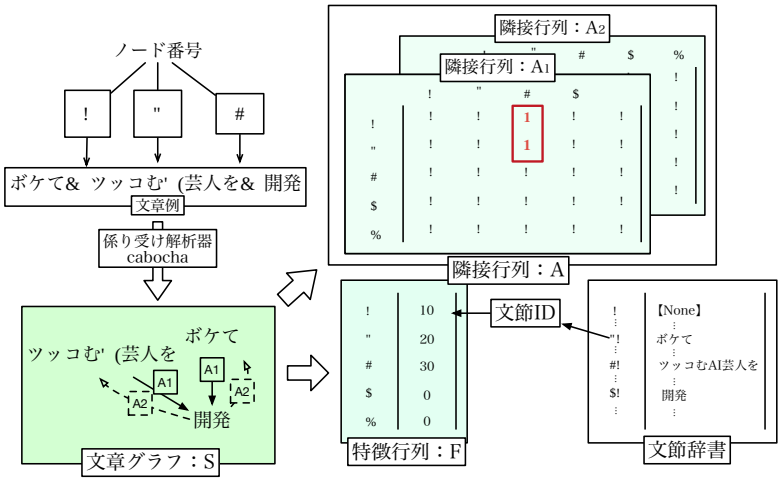

図４　隣接行列と特徴行列の獲得過程

【学習データ】

・H7N9 ヒトヒト感染疑い相次ぐ
・点眼のあと目パチパチはダメ

【ルールベースによるかさ増し例】

・ニホンオオカミのナメクジ多い岡山大
・けつえきリプログラミングの説明いでんしを多い

【SeqGAN(単語) によるかさ増し：単語の繰り返しが見られる例】

・100 万年前のマネキン巻き貝化石化石
・皮むける雷鳥最長油油油搭載油を水銀に

【TextGAN(単語) によるかさ増し：記号の繰り返しが見られる例】

・無比暑熱のの「「「「新種
・氏病のの「「「「「「「「「「

【SeqGAN(文節) によるかさ増し：同一の文章が見られる例】

・ウイルス作成咎に疑問のそれ
・ウイルス作成咎に疑問のそれ

【TextGAN(文節) によるかさ増し：単語の繰り返しが見られる例】

・3 歳男児が上限上限上限上限上限上限
・城郭から城郭から両腕で新種肉食恐竜近縁亡霊りゅうし痕跡?

【提案手法によるかさ増し：比較的自然な生成が行われた例】

・脳波で涙ほっとした山中氏ら問題問題解決能力分野で太陽光発電，交渉バイオ燃料目標羽毛恐竜全身の
・使って目覚め終末期苦痛緩和望む移送の絶望 ISS，仕組み解明名大大地震太平洋’ 沿岸で NASA 新宇宙服

図５　学習データと文章生成結果

を用いた生成手法では変化の大きい生成が行えるが，不自然な文章も生成されやすくなる．提案手法と他の GAN を用いた手法の生成結果を比較すると，文節単位の学習を行った SeqGAN では同一の文章の繰り返しが，TextGAN では単語の繰り返しが見られた．また単語単位で学習を行った SeqGAN と TextGAN の生成では単語の繰り返しが見られる．単語の繰り返しは違和感のある文章を生成する要因であり，文章生成においてはノイズになりうる．そこで，生成手法毎の単語の繰り返し数から自然な文章らしさを評価する．

表 4 は学習データ，手法毎の生成文章それぞれ 14,800 件に

表３　実験に用いたパラメータ

モデル	epoch 数	バッチサイズ	λ	次元数
word2vec	5	32	-	300
TextGAN	180	128	0.2	32
SeqGAN	180	128	0.2	32
提案手法	30	32	0.8	8

表４　単語の繰り返し回数比較

繰り返しの有無	無し	有り
学習データ	14792	8
ルールベース	14726	74
SeqGAN(単語)	14619	181
TextGAN(単語)	11924	2876
SeqGAN(文節)	14402	598
TextGAN(文節)	13416	1584
提案手法	14748	52

対し，単語の繰り返しが見られる文章数を表したものである．学習データに見られる繰り返しは「パチパチ」や「○○」などの口語に近いものであるが，生成文章では単語や記号の繰り返しが多く見られた．最も単語の繰り返し数が少なかったのは提案手法であり，これはグラフ構造を用いた事により比較的自然な文章が生成できたと言える．次に，生成文章と学習データの類似度を頻出単語の一致率からみる．学習データ内の頻出単語上位 n 件と生成文章内の頻出単語上位 n 件を比較し，学習データの高頻度語が生成データの高頻度語にどれだけ含まれるかを見る．単語は名詞に限定している．

図 6 は学習データの高頻度語上位件数 n を 10 から 1000 まで 10 刻みで変化させた時の一致率を表したグラフである．縦軸は一致率，横軸は頻出単語数を表し，一致率が高いほど学習データの特徴を獲得出来たといえる，ルールベース，単語単位で学習した SeqGAN，TextGAN，文節単位で学習した SeqGAN，TextGAN，提案手法の 6 つの生成手法を用い，学習データとの一致率を比較すると，全ての場合で提案手法が最も高い一致率を示しており，学習データの特徴を十分に獲得出来た事がわかる．以上の結果から提案手法は従来の生成手法に比

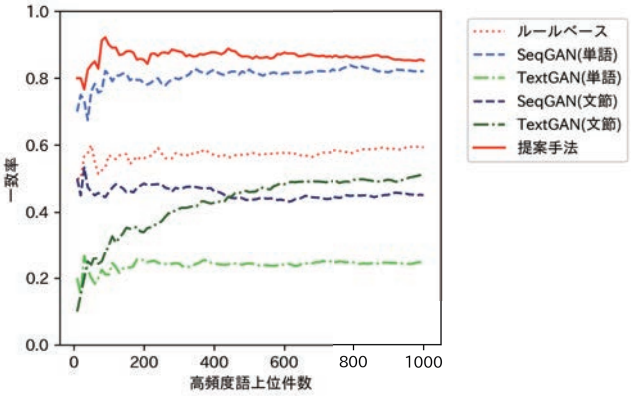

図 6　学習データと生成文章の頻出単語の一致率

べ，違和感の少ない文章を学習データの特徴を獲得した上で生成出来ている事がわかった．次に，かさ増し手法として用いる場合には単語ではなく文章単位の類似度が求められるため，文章の類似度比較による評価を行う．

4.2　類似度比較実験

かさ増しは元データの特徴を獲得しつつ異なるデータを多様に生成することが求められる．元データの特徴を獲得できない，あるいはデータの多様性が低い場合，かさ増し手法として分類問題に用いると精度の低下などの悪影響が発生しうる [7]．そのため，かさ増しの評価行うためには元データとの類似性と生成データの多様性の 2 つを評価する必要がある．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を評価する実験を行うことで，提案手法のかさ増し手法としての評価行う．学習データと生成データの類似度が高いほど学習データの特徴を獲得出来たと評価し，生成データ間の類似度が低いほど多様性があると評価される．類似度計算は BLEU スコアと Doc2vec [22] を用いて獲得した文章ベクトル間の cos 類似度を用いる．BLEU スコアが高ければ語彙の類似性が高いという事であり，cos 類似度が高ければ文脈としての類似性が高いと評価することができる．

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，10,000 件を生成データとして生成した．類似度計算は学習データからランダムに 1,000 件，生成データからランダムに 1,000 件の文章を取り出し，それぞれについて計算した類似度の平均を評価値としている．ランダムによる偏りを防ぐため，10 回実験を行い箱ひげ図を作成した．

4.2.1　学習データとの類似度比較実験

図 7 は学習データと生成データの類似度を BLEU スコアで計算した結果を表し，図 8 は cos 類似度を用いて計算した結果を箱ひげ図を用いて表したものである．類似度が高いほど元の文章に近い文章が生成できているといえ，学習データの特徴を獲得できたといえる．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．

提案手法を他の手法と比較すると，カテゴリ，評価指標によらず高い類似度を示していることがわかる．BLEU スコアでは

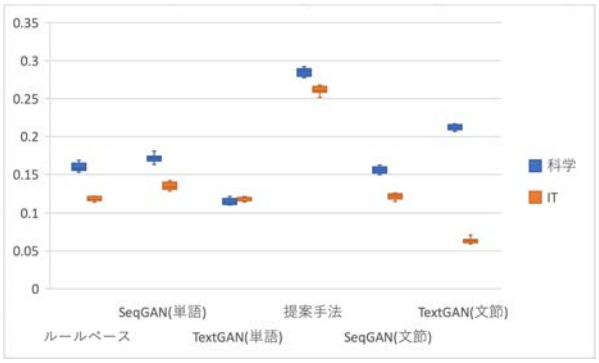

図 7　学習データとの類似度比較実験（BLEU）

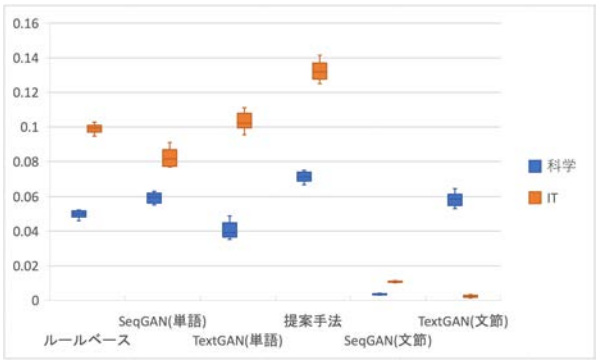

図 8　学習データとの類似度比較実験（cos）

提案手法が他の生成手法と比べ，IT カテゴリと科学カテゴリの両方で高い類似度を示している．ここから単語単位で見た学習データの特徴を獲得できたといえる．BLEU スコアは分類を行う上でも重要な特徴とされており [23]，有効なかさ増しが行えたことが示唆された．提案手法では単語ではなく文節区切りを用いた生成を行なっているため，BELU スコアが類似しやすい可能性がある．そこで SeqGAN と TextGAN について，単語区切りと文節区切りのそれぞれで学習を行った結果を比較すると，どちらも文節区切りでは精度が低下する傾向が見られた．これは文節区切りでの学習では学習データの特徴を十分に獲得出来ないことを示し，MolGAN を用いたことにより文章全体の構造を有効に獲得できた結果といえる．

4.2.2　生成データ内の類似度比較実験

図 9 は生成データ内の類似度を BLEU スコアで計算したもので，図 10 は cos 類似度を用いて計算した結果を箱ひげ図で表している．類似度は数値が低いほど多様性のある文章が生成できていると評価する．コーパス毎の多様性を見るため，文書ベクトルはそれぞれのコーパス毎に学習している．ベースラインとして学習データ内の類似度を算出した．これにより実際のデータセットの多様性との比較を行うことが出来る．

まず学習データの類似度からカテゴリごとに異なる特徴がある事がわかる．IT カテゴリは科学カテゴリより学習データ内の類似度が高い．BLEU スコアの類似度が高いことから，IT カテゴリでは同じ単語が出やすい傾向が示唆され，cos 類似度から IT カテゴリでは定型文のような類似した文章が多い傾向

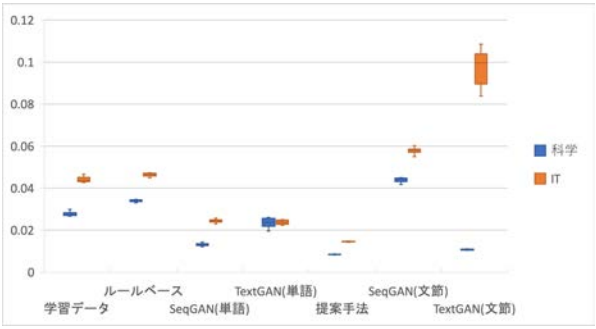

図９　生成データ内の類似度比較実験（BLEU）

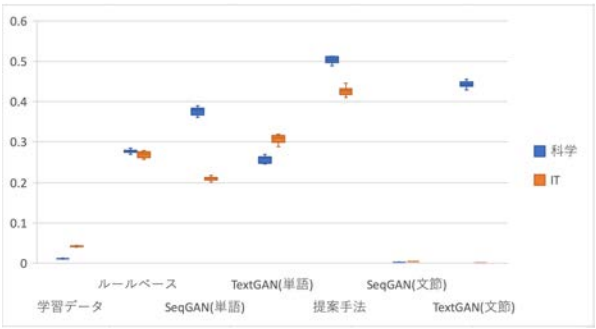

図 10　生成データ内の類似度比較実験（cos）

表５　不均衡データ

カテゴリ名	適合率	再現率	F 値
国際	0.57	0.35	0.43
国内	0.33	0.36	0.35
経済	0.36	0.12	0.18
エンタメ	0.31	0.29	0.30
スポーツ	0.32	0.60	0.42
IT	**0.27**	**0.05**	**0.08**
科学	**0.37**	**0.07**	**0.11**
地域	0.39	0.47	0.42
F 値の標準偏差		**0.1442**	
全体精度		**0.3541**	

表６　均衡データ

カテゴリ名	適合率	再現率	F 値
国際	0.56	0.42	0.48
国内	0.32	0.22	0.26
経済	0.22	0.34	0.27
エンタメ	0.24	0.38	0.30
スポーツ	0.33	0.25	0.29
IT	**0.47**	**0.27**	**0.34**
科学	**0.22**	**0.34**	**0.27**
地域	0.51	0.33	0.40
F 値の標準偏差		**0.0774**	
全体精度		**0.3181**	

が示唆された．

次に生成手法の類似度を見ると，BLEU スコアでは高い多様性が見られるが，cos 類似度でみたときの多様性は学習データよりも低い．これは単語単位で見た場合は多様だが，文章単位で見た場合文章がある程度類似している可能性が考えられる．多様性の最も高い文節区切りを用いた SeqGAN を見ると，図 7，図 8 より特徴の獲得が不十分である．文節区切りを用いた TextGAN でも同様に，特徴を獲得できなかった IT カテゴリにおいて cos 類似度での多様性が高い．一方，単語を用いた生成手法では，cos 類似度での多様性が低い傾向がある．このことから特徴の獲得が不十分な場合，ランダムノイズのような文章を生成してしまう可能性がある．よって文章の特徴獲得により，ある程度多様性は低下するといえる．

提案手法は BLEU スコアでは多様な生成であることを示し，cos 類似度では多様性の低い生成結果であることを示している．これは特徴を獲得した結果多様性が低下したと考えられ，カテゴリ毎の特徴的な文章の構造をある程度獲得し生成していると考えられる．以上のことから，提案手法は他の手法に比べ学習データに近い多様な生成が行えていることが示唆された．そこで，実際にかさ増しとして用いた場合の影響を分類実験により評価する．

4.3　分類実験

分類実験によって提案手法によるかさ増しが，カテゴリの特徴を獲得しているかを評価する．かさ増し手法が分類に有効な特徴を獲得していた場合，かさ増しデータにより精度が向上する．実験に用いるデータセットは表 1 に示す，かさ増し無しの不均衡なデータ，均衡データ，かさ増し不均衡データ，かさ増し均衡データの 4 種類である．

均衡データは，IT，科学カテゴリに合わせて 14,000 件にデータ数を揃えたものであり，不均衡データを分類する際に用いられるダウンサンプリングを行ったデータセットである．かさ増し不均衡データは IT，科学カテゴリに対して 30,000 件になるように提案手法を用いてかさ増しを行ったデータセットであり，他のカテゴリについては不均衡のままである．かさ増し均衡データはかさ増ししたカテゴリに合わせて全カテゴリのデータ量を 30,000 件に均衡化したデータセットである．

分類は TF-IDF を用いた 100 次元のベクトルに対しランダムフォレストを用いる．学習データを全体の 8 割，テストデータを全体の 2 割とする．かさ増しデータはテストに含まない．

4.3.1　実験設定

不均衡データを用いて実験した結果を表 5 に，均衡データを用いた結果を表 6 に，かさ増しデータを用いたものを表 7 に，かさ増し均衡データを用いた実験結果を表 8 に示す．各データに対する分類性能を全体精度で評価する．しかし，全体精度は正解数の割合を表すため偏った分類を行なった場合と，安定した分類を行なった場合の区別が難しい．このため，評価指標として各カテゴリの適合率，再現率，F 値と F 値から計算した標準偏差を用いる．適合率を見ることで分類の正確性，再現率を見ることで分類数の傾向をそれぞれ見る事ができる．F 値は適合率と再現率から求められ，両方を加味したカテゴリの精度を見る事ができる．F 値の標準偏差は分類能力におけるカテゴリ

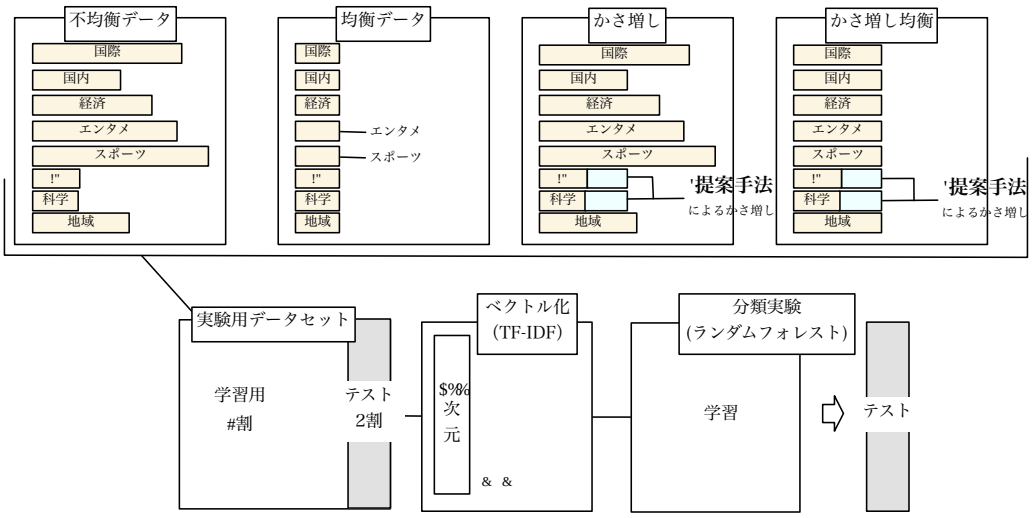

図 11　分類実験概要

表 7　かさ増し不均衡データでの実験

カテゴリ名	適合率	再現率	F 値
国際	0.47	0.37	0.42
国内	0.29	0.33	0.31
経済	0.34	0.15	0.21
エンタメ	0.30	0.26	0.28
スポーツ	0.29	0.63	0.40
IT	**0.75**	**0.25**	**0.38**
科学	**0.38**	**0.11**	**0.17**
地域	0.44	0.43	0.43
F 値の標準偏差		**0.0987**	
全体精度		**0.3478**	

表 8　かさ増し均衡データでの実験

カテゴリ名	適合率	再現率	F 値
国際	0.47	0.42	0.44
国内	0.38	0.35	0.37
経済	0.26	0.30	0.28
エンタメ	0.22	0.41	0.29
スポーツ	0.28	0.41	0.33
IT	**0.77**	**0.32**	**0.45**
科学	**0.44**	**0.16**	**0.24**
地域	0.25	0.24	0.25
F 値の標準偏差		**0.0817**	
全体精度		**0.3271**	

の依存性を評価する事ができる．これらの指標を用いて，かさ増しによる効果をデータセット間の比較により評価する．

4.3.2　実験結果と考察

まず表 5 の不均衡データを見ると，全体精度が最も高いが標準偏差も最も大きい．これはカテゴリ毎の分類が不安定である事を示し，全体精度の高さは実際の分類能力を表していないと考えられる．次に表 6 の均衡データを見ると，不均衡データと比較し標準偏差が改善されている事がわかる．全体精度が低下しているが各カテゴリが安定して分類された結果であるため，本来の分類性能を示しているといえる．表 7 のかさ増し不均衡データ，表 8 のかさ増し均衡データも均衡データに近い標準偏差であることから，不均衡データに対しては何らかのサンプリング手法が必要であることが分かる．また，均衡データと，かさ増し不均衡データ，かさ増し均衡データを比較すると，かさ増しによって全体精度が向上していることがわかる．これはかさ増しにより増加したデータが分類能力の向上に寄与した結果であり，類似度比較実験で見られた多様性の低下は特徴獲得によるものであったと結論付けられる．よって，提案手法によるかさ増しはデータの不均衡さの解消に貢献し有効な手法であると考えられる．かさ増し不均衡データ，かさ増し均衡データを比較すると，表 7 のかさ増し不均衡データでは全体精度が高く，表 8 のかさ増し均衡データでは標準偏差が改善されていることが分かる．全体精度はデータ量に依存し，標準偏差はデータ量の均衡さに依存するといえ，問題設定に応じて適切なサンプリングを選択する必要性を示している．

かさ増しによる分類結果の変化から，かさ増しによる影響を考察する．表 6 の均衡データ，表 8 のかさ増し均衡データを見るとかさ増し対象では無い，国内カテゴリの F 値が向上していることが分かる．また，地域カテゴリでは F 値の低下が見られる．これはかさ増しによって増加した特徴が全体に影響を及ぼしていると考えられ，かさ増しの効果を分類前に予測する事が困難である事が分かる．

かさ増し対象である IT カテゴリと科学カテゴリに注目すると，IT カテゴリは不均衡データ時の 8% が最小値であり，かさ増し均衡データ時の 48% が最大値である．科学カテゴリは不均衡データ時の 11% が最小値であり，均衡データ時の 27% が最大値になった．どちらもはサンプリング手法により F 値が向上しているが，科学カテゴリはかさ増しでの精度向上率が小さい．この向上率の差は提案手法による特徴獲得能力にカテゴリ差があることを示している．表 6 の均衡データと表 8 のかさ増し均衡データを比較すると，特徴を十分に獲得出来た IT

カテゴリでは適合率，再現率共に向上しており，非常に有効なかさ増しが行えている事がわかる．一方で，十分に特徴を獲得出来なかった科学カテゴリでは適合率が向上したが再現率が低下し，F 値の低下につながっている．これは過学習の傾向を表わしており，生成された特徴の多様性が低いことが原因だと考えられる．今回全てのデータ量が最大になる，スポーツカテゴリに合わせて各カテゴリのデータ量を増加させ均衡化する実験を行わなかったのは，この多様性の低下によるものである．多様性の低下の原因として，新たな語彙を獲得できないことがある．生成手法は学習データの語彙空間に大きく依存する．より汎用的なかさ増しを行うためには新たな語彙を獲得する工夫が必要である．また，表 2 より，文節区切りの精度が高い IT カテゴリの方が科学カテゴリよりも精度の向上が見られることから，文節区切りの精度の向上が特徴の獲得に寄与することが示唆される．

5. まとめ

本論文では CaboCha により生成したグラフ構造を加味した文章生成を行うために GAN に R-GCN を採用したモデルを提案した．入力に用いる隣接行列はグラフの方向，特徴行列は文節 ID を用い，生成結果を分析するために生成実験，類似度比較実験，分類実験の 3 つを行った．生成実験では提案手法が比較的特徴を含む自然な文章を生成している事が分かった．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多様な生成が行えていることが示された．分類実験の結果，分類に有効な特徴をかさ増し出来ていることが確認できた．同時にかさ増しの効果が表れにくいカテゴリの存在も確認され，より多様な特徴を捉えられるかさ増しシステムの提案が必要になる．結論として，提案手法を用いる事でデータの総量を確保したかさ増しを行う事が可能であり，データの不均衡さの解消に有用であるといえる．また，文節区切りの精度向上が学習データの特徴を捉えたかさ増し手法に寄与することが示唆された．

今回用いたかさ増し手法は学習データとの類似度や多様性を評価したが，これらが分類精度にどの程度寄与するかは解明されていない今後の課題として，文章分類問題における学習データの特徴の解析を解析し，精度向上に寄与するかさ増し手法の提案があげられる．

参 考 文 献

[1] H. He and E. A. Garcia: “Learning from imbalanced data,” *IEEE Trans. on Knowledge and Data Engineering,* Vol.21, No.9, pp. 1263-1284, 2009.

[2] I. Goodfellow et al.: “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems,* 2014.

[3] G. Mariani et al.: “BAGAN: Data Augmentation with Balancing GAN,” arXiv preprint, arXiv:1803.09655, 2018.

[4] C. Coulombe: “Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs,” arXiv preprint, arXiv:1812.04718, 2018.

[5] S. Kobayashi: “Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations,” arXiv preprint, arXiv:1805.06201, 2018.

[6] CaboCha/南瓜, Yet Another Japanese Dependency Structure Analyzer: https://taku910.github.io/cabocha/

[7] 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.

[8] Word2vec: https://code.google.com/p/word2vec/ [accessed Oct. 4, 2018]

[9] H. Isahara et al: “Development of the Japanese WordNet,” *Proc. of the Sixth Int. Conf. on Language Resources and Evaluation*, 2008.

[10] L. Yu et al.: “SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient,” *AAAI*. 2017.

[11] Y. Zhu et al.: “Texygen: A Benchmarking Platform for Text Generation Models,” arXiv preprint, arXiv:1802.01886, 2018.

[12] M. J. Kusner and J. M. Hernández-Lobato: “Gans for Sequences of Discrete Elements with the Gumbel-softmax Distribution,” arXiv preprint, arXiv:1611.04051, 2016.

[13] T. Che et al.: “Maximum-Likelihood Augmented Discrete Generative Adversarial Networks,” arXiv preprint, arXiv:1702.07983, 2017.

[14] K. Lin et al.: “Adversarial Ranking for Language Generation,” arXiv preprint, arXiv:1705.11001, 2017.

[15] J. Guo et al.: “Long Text Generation via Adversarial Training with Leaked Information,” arXiv preprint, arXiv:1709.08624, 2018.

[16] Y. Zhang et al.: “Adversarial Feature Matching for Text Generation,” arXiv preprint, arXiv:1706.03850, 2017.

[17] K. Papineni et al.: “BLEU: a Method for Automatic Evaluation of Machine Translation,” *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318, 2002.

[18] M. Schlichtkrull et al.: “Modeling Relational Data with Graph Convolutional Networks,” arXiv preprint, arXiv:1703.06103, 2017.

[19] N. D. Cao et al.: “MolGAN: An Implicit Generative Model for Small Molecular Graphs,” arXiv preprint, arXiv:1805.11973, 2018.

[20] M. Simonovsky et al.: “GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders,” arXiv preprint, arXiv:1802.03480, 2018.

[21] Yahoo!ニュース: https://news.yahoo.co.jp [accessed Oct. 4, 2018]

[22] Doc2vec: https://radimrehurek.com/gensim/models/doc2vec.html

[23] D. R. Amancio et al.: “Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts,” *Physica A: Statistical Mechanics and its Applications*, Vol.391, No.18, pp. 4406-4419, 2012.

（2019 年 3 月 21 日　受付）
（2019 年 9 月 2 日　採録）

［問い合わせ先］
〒903-0213　沖縄県西原町千原 1 番地
琉球大学
澤崎　夏希
TEL: 098-895-8589
E-mail: k178577@ie.u-ryukyu.ac.jp

著 者 紹 介

澤崎（さわさき）　夏希（なつき）［非会員］

2017 年琉球大学工学部情報工学科卒業．2019 年琉球大学理工学研究科情報工学専攻修了．人工知能学会会員．

遠藤 聡志［正会員］

1990 年北海道大学大学院工学研究科電気工学 専攻修士課程修了．同年，北海道大学工学部助手．1995 年琉球大学工学部情報工学科講師．1996 年同助教授．2004 年同教授. 複雑系工学に関する研究に従事．情報処理学会，人工知能学会，計測自動制御学会各会員．博士（工学）．

山田 孝治［非会員］

1995 年北海道大学大学院工学研究科情報工学専攻修了．博士（工学）．同年，琉球大学工学部情報工学科助手，1996 年同講師，1999 年同准教授.2014 年同教授．マルチエージェント，知能ロボットに関する研究に従事．情報処理学会，機械学会，ロボット学会，人工知能学会各会員．

當間 愛晃［非会員］

2003 年琉球大学大学院理工学研究科総合知能工学専攻（博士後期課程）修了．博士（工学）．2004 年琉球大学工学部情報工学科助手．2007 年同大学助教．2015 年同大学准教授．複雑系工学，データ/テキスト/Web マイニング，人工知能に従事．自然言語処理学会，日本認知科学会，人工知能学会各会員．

赤嶺 有平［非会員］

2004 年琉球大学大学院理工学研究科博士課程総合知能工学専攻修了．博士（工学）．同年，日本学術振興会特別研究員．2006 年琉球大学工学部情報工学科助手．2007 年から同助教．交通システム，複合現実感の研究に従事．地理情報システム学会，人工知能学会各会員．

Sentence Generation Method by Extension of MolGAN Using Sentence Graph

by

Natsuki SAWASAKI, Satoshi ENDO, Naruaki TOMA, Koji YAMADA, and Yuhei AKAMINE

Abstract:

Deep learning solves many classification problems. However, it is difficult to solve problems with imbalanced data. Therefore, the data volume is increased for the purpose of balancing. This is called data augmentation. Generally, the method of image data augmentation uses noise addition, rotation, and the like. Recently, images are generated using the generative adversary network: GAN. However, data augmentation methods are difficult in natural language processing. In addition, manual data augmentation is burdensome and requires mechanical methods. Mechanical text augmentation is more difficult than images. Because it is difficult to analyze the feature of sentences. This paper proposes a sentence generation method by machine learning focusing on graph information. The graph information obtained by CaboCha is processed by graph Convolution. The proposed GAN was used to generate sentences, and then three experiments were performed to evaluate its effectiveness.

Keywords: text data augumantation, unblanced data, GAN

Contact Address: **Natsuki SAWASAKI**
University of Ryukyus
1 Senbaru, Nishihara-cho, Okinawa 903-0213, Japan
E-mail: k178577@ie.u-ryukyu.ac.jp