5V-05

SeqGANを用いたマイクロブログの文章自動生成に関する研究

三村亮† 上田芳弘† 坂本一磨†
公立小松大学生産システム科学部†

１．はじめに

SNS (Social Networking Service) の普及に伴い，膨大な一般ユーザの情報がネットワーク上に蓄積されている．この情報を活用し，ユーザのニーズや属性などを推定，分類することにより，マーケティングに応用する取り組み[1]が行われている．しかし，深層学習を用いて分類問題を解く際，各カテゴリのデータ数が不均衡な場合，学習が困難な課題がある．その課題を解決するため，データを事前にサンプリングし，データ量の不均衡さを解消する方法が用いられる．サンプリング手法は，多数派カテゴリのデータ量を減らすアンダーサンプリング，少数派カテゴリのデータ量を増やすオーバーサンプリングに大別される．アンダーサンプリングを用いた場合，学習に必要な情報も削減するため，特徴量を学習しきれない課題が発生する．一方，オーバーサンプリングはデータの損失が無いため，十分な学習が期待できる．オーバーサンプリングの既存研究[2]では，ニュース記事をデータセットに用いた手法が提案されている．しかし，ニュース記事は画一的な文体で書かれており，様々な文体で表現されるマイクロブログの文章には適応が困難である．そこで，本論文では，マイクロブログ上の文章の属性をSeqGAN[3]を用いて学習，推定し，ユーザ属性に基づいた文章の生成を行い，オーバーサンプリングを行う手法を提案する．

２．提案手法

本システムの概要を図1に示す．本提案手法は形態素解析と文章生成モデル構築機能，文章生成機能で構成される．本研究では，マイクロブログの1つであるTwitter上の投稿を対象とした．実験データの収集方法は，Twitter のプロフィールから任意の文字列を検索できる，ツイプロ[4]を用いて，プロフィール欄に性別を明記しているユーザの投稿を無作為に収集した．

２．１　形態素解析

本機能では，形態素解析エンジンであるMeCabを用いて入力データの文章の形態素解析を行い，分かち書きを行う．形態素とは，文を品詞単位に分割したものである．また，文中のノイズとなるURLや記号などは除去する．

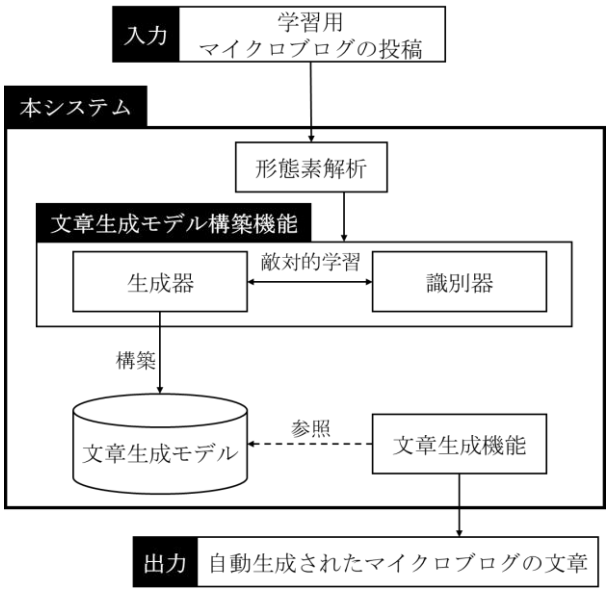

図1　本システムの概要

２．２　文章生成モデル構築機能

本機能では，SeqGAN を用いて，文章生成モデルを構築する．SeqGAN は，生成器と識別器を用意し，それらを敵対的に学習させる手法であるGAN を系列データに対応できるように発展させたシステムである．生成器にはLSTMを使用し，識別器には，CNN を使用する．生成器では，モンテカルロ探索を用いて，文章の中間状態と行動選択を識別器で評価し，それらを報酬とした強化学習を行う．また，識別器は生成された文章とデータセットとして用いた文章を正しく識別するように学習を行い，生成器は識別器に識別されない文章の生成に向けた学習を行う．なお，これらの学習は交互に行う．

入力データは，形態素解析を行い，分かち書きしたテキストデータである．入力データを読み込む際，形態素ごとに単語 ID を割り当てた辞書が作成され，その辞書を基に形態素と単語 ID の変換を行う．また，単語 ID は行列を用いたベクトル表現に変換する．

２．３　文章生成機能

本機能では，構築した文章生成モデルを参照し，自動で文章を生成する．

Research for Automatic Text Generation for Microblogging Using SeqGAN
† Aki Mimura, Yoshihiro Ueda, Kazuma Sakamoto
Faculty of Production Systems Engineering and Sciences, Komatsu University

３．評価実験

本実験では，文章生成実験と評価実験の二つの実験を行った．

３．１　文章生成実験

本実験では，投稿者の性別が判別している Twitter 上の投稿をデータセットに用いて文章の生成を行う．データセットは男女それぞれ 10,000 件の投稿を用いた．

３．２　文章生成結果と考察

提案手法による生成例を図 2 に示す．この生成例の他に文法的に正しいが，単語ごとの脈略が無いワードサラダと呼ばれる文章が多く見られた．しかし，ワードサラダの文章であっても，属性推定には有用であると判断したため，本研究ではそのまま処理する．

３．３　評価実験

本提案手法を用いて生成された文章のオーバーサンプリングでの有効性を評価するために．Bi-LSTM を用いて男女の 2 値分類精度を比較する．テストデータは学習に使用したデータとは別の Twitter 上の投稿，男女それぞれ 1,000 件とした．学習に用いるデータセットを表 1 に示す．データセット 1 は提案手法の学習に用いた Twitter 上のユーザ投稿のみ 4,000 件である．データセット 2 は本提案手法で生成した文章のみ 4,000 件である．データセット 3 はデータセット 1，データセット 2 をそれぞれ 2,000 件ずつ，無作為に抽出し，混合させたものである．なお，全てのデータセットにおいて，男女の割合は等しくなっている．

３．４　結果と考察

評価実験の結果を表 2 に示す．指標となるデータセット 1 の正解率が 0.542 と低く，分類器である Bi-LSTM が十分に学習できていないと考える．その原因は，データセットのデータ数が少なかったために，過学習が発生している可能性があると考えられ，データを増加し，検証する必要性がある．その後に，今回の実験結果と同様に，データセット 1 と比較してデータセット 2 と 3 の正解率や適合率，再現率，F 値が大きく低下することがないか確認する予定である．

４．おわりに

本研究では，SeqGAN を用いて文章を生成し，オーバーサンプリングを行う手法を提案した．今回は男女の 2 値の属性に対してのみ学習を行ったが，今後は，データ数を増加し，3 値以上の属性に発展することを検討する．また，提案手法の評価が不十分であった．そのため，今後は評価方法の見直しを行う．

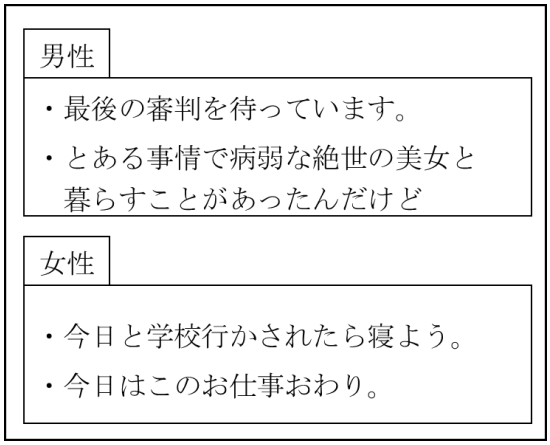

図 2　提案手法による生成例

表 1　評価実験に用いるデータセット

	データセット 1	データセット 2	データセット 3
Twitter 投稿	4,000 件	0 件	2,000 件
提案手法	0 件	4,000 件	2,000 件

表 2　評価実験の結果

	データセット 1		データセット 2		データセット 3	
	男性	女性	男性	女性	男性	女性
正解率	0.542		0.511		0.508	
適合率	0.554	0.535	0.511	0.511	0.510	0.508
再現率	0.437	0.647	0.501	0.521	0.437	0.580
F 値	0.489	0.586	0.506	0.516	0.471	0.541

参考文献

[1] 池田和史，服部元，松本一則，小野智弘，東野輝夫：マーケット分析のための Twitter 投稿者プロフィール推定手法，情報処理学会論文誌コンシューマ・デバイス＆システム（CDS），Vol.2，No.1，pp.82-93，2012.

[2] 澤崎，遠藤，當間，山田，赤嶺：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，知能と情報（日本知能情報ファジィ学会誌），Vol.32, No.2,pp.668-677, 2020.

[3] SeqGAN: Sequence generative adversarial nets with policy gradient, L. Yu, W. Zhang, J. Wang, and Y. Yu, AAAI, pp.2852-2858, 2017

[4] S21G 社：ツイプロ，入手先<http://twpro.jp/>（参照 2022-1-6）