変換規則と文章の分散表現に基づく自動文章生成手法の提案

福田 清人　　森 直樹　　松本 啓之亮
大阪府立大学 工学研究科
{fukuda@ss., mori@, matsu@}.cs.osakafu-u.ac.jp

1 はじめに

近年, 人の感性に基づく物語を計算機によって自動生成させる試みが注目を集めている. 小説や漫画の様な物語はストーリーと表現媒体という 2 つの要素に分解される. ストーリーは物語の内容であり, 表現媒体はストーリーを表現するための媒体である.

筆者らはこれまで物語のストーリーの自動生成に焦点を当て, シミュレーションに基づいたストーリーの半自動生成手法 [1] を提案してきた. その結果, 現状では生成に利用したストーリーとは異なる, 短いストーリーの半自動生成が可能となりつつある. しかしながら, 生成されたストーリーから小説を生成するためには大部分を人手に頼らざるを得ない. これは, 既存の文章を用いずに新たな文章を自動生成することが非常に困難な課題である.

既存の文章に基づく文の自動生成に関する研究はこれまでにいくつか報告されており, 文法的には違和感のない文章を生成することが可能な手法が提案されている. しかしながら, 意味的にも違和感のない文章を生成することが困難である点や, 文章の妥当性を定量的に評価していないという点が問題点として挙げられる.

一方, 近年機械学習に基づいて獲得した単語や文の分散表現を用いることで, 単語や文間の意味的な演算が可能であることが報告されている. そこで, 分散表現を用いて文章の意味的な要素を考慮した上での定量評価が可能ではないかと考えた.

以上の観点から, ジャンルに依らないストーリーに基づいて任意のジャンルの小説を自動生成することを本研究の最終目標とする. その第一歩として, 本研究では文章の分散表現と変換規則を用いたジャンル風文章の自動生成手法を提案する. また, 提案手法で生成された文章の主観評価をすることで, 提案手法の有用性を確認した.

2 関連研究

本研究と関連が強い研究として, 既存の文章を用いた文の自動生成に関する研究や文章間の類似度の計算に関する研究が挙げられる. それぞれについて以下で述べる.

2.1 既存文を用いた文の自動生成

模倣元となる既存の文章の従属節を, 置換に利用する文章の従属節で置換することで新たな文章を自動生成する手法 [2] が提案されている. この手法では置換元と置換先の従属節の組を決定し, 適切な形に変形した上で置換する. 従属節の単位で置換することで, 文法的に正しい文章が生成できることが報告されている. しかしながら, 無作為に従属節の組を決定しているため, 意味的に成立していない文章が生成される可能性が高いという問題点が存在する.

また, 既存文に対して事前に用意した変換規則を適用することで, 文に特定の属性や特徴を付与する手法 [3] が報告されている. 文に対して, 適用できる候補の中から変換規則を選択し, 確率的に置換することで文に属性や特徴を付与する. その結果, 日本語として違和感のある出力を少なく抑えつつ, 人間がある程度判別可能な特徴を付与できることが報告されている. しかしながら, 変換規則ができる箇所が限定されており, 意味の異なる文章を生成することができないという問題点がある.

2.2 文章間の類似度計算

近年, 単語間の意味的な演算ができるとして, 単語の分散表現が大きな注目を集めている. 分散表現を得る代表的な手法として, Word2Vec [4] がある. Word2Vec では, テキストに出現する各単語を周辺の単語から予測する単語予測タスクを, ニューラルネットワークに

より学習することで, 単語に対する概念ベクトルを獲得する. 得られた分散表現により単語を意味空間上の 1 点に対応させることができ, 単語間の意味的な類似度を計算することが可能となる.

Word2Vec を文章に拡張したものが Doc2Vec [5] である. Doc2Vec は文章から単語ベクトルとパラグラフベクトルを生成し, 平均化と結合, 推定を経て文章の分散表現を構築する. 文章の分散表現では, 単語の分散表現同様に文章間での意味を考慮した類似度が計算できると考えられる.

3 提案手法

本研究では, 事前に自動生成した変換規則を, 既存の文章に適用して新たな文章を生成し, 目標となるジャンルの文章との類似度を分散表現により計算することで, 任意の文章から特定のジャンル特徴を有した文章を自動生成する手法を提案する. 以下に提案手法の概要を示す. 各ステップの詳細については 3.1 〜 3.3 節で述べる.

Step.1 変換規則の自動生成
変換したいジャンルの作品を含むいくつかの作品を用いて変換規則を獲得する.

Step.2 Doc2Vec のモデル構築
複数の作品を用いて, Doc2Vec のモデルを学習する.

Step.3 ジャンル特徴を有した文章の生成
任意の文章に対して, 変換規則の適用と分散表現による類似度計算を繰返すことで, 特定のジャンル風な文章を自動生成する.

3.1 変換規則の自動生成

提案手法で用いる変換規則は, 対象となる文章中の文節を異なる文節と置換するための規則であると定義する. 変換規則は既存のテキストデータから機械的に自動生成される. 以下に変換規則の生成アルゴリズムを示す.

1. 変換規則生成用のテキストデータを句点を基準として分解し, 文集合 C とする. ただし, 文章の先頭と末尾に括弧が存在する場合, 会話文とみなし, 括弧内の文章全体を 1 文とみなす.
2. C 中の i 番目の文 $c_i(i = 1, 2, \cdots, |C|)$ を抽出する.
3. 文 c_i に対して構文解析し, 得られた文節集合を P_i とする.
4. 文節集合 P_i 中の j 番目の文節 $p_{i,j}(j = 1, 2, \cdots, |P_i|)$ を抽出する.
5. 文節 $p_{i,j}$ に対して, 直前の文節 $p_{i,j-1}$ および $p_{i,j}$ の末尾情報 $e_{i,j-1}, e_{i,j}$ を取得する. また, $p_{i,j-1}$ の先頭の形態素における品詞情報 $f_{i,j}$ を取得する. ここで, 文節の末尾情報は文節における末尾の形態素の品詞情報と表層表現を組合せたものであり, $e_{0,j}$ には文頭であるという情報が入る.
6. 4 で取得した $e_{i-1,j}, e_{i,j}, f_{i,j}$ および文節 $p_{i,j}$ を要素に持つ集合 $R = \{e_{i-1,j}, e_{i,j}, f_{i,j}, p_{i,j}\}$ を変換規則として取得する. ここで, $e_{i-1,j}, e_{i,j}$ は変換の適用条件, $f_{i,j}$ は変換元の条件, $p_{i,j}$ は変換先となる.

表 1 に変換規則の例を示す.

3.2 Doc2Vec のモデル学習

提案手法では, 生成される文章に対して目標となるジャンルの文章との意味的な類似度を計算するために Doc2Vec を利用する. 同一ジャンルの文章ばかりを学習データとして与えてしまうと, ジャンルの特徴を学習することができないため, 今回は学習データに, 目標となるジャンル以外の文章も含めた. Doc2Vec には Python の gensim モジュールを利用する.

3.3 ジャンル風文章の生成

提案手法では, 基準となる文章に対して, 変換規則を適用することで新たな文章を生成する. 生成された文書は, 目標となるジャンルの文章との分散表現での類似度によって評価される. 上述した 2 つの操作を繰返すことで, ジャンル風の文章を自動生成する. 以下にジャンル特徴を有した文章の生成アルゴリズムを示す.

1. 目標となるジャンルの文章を改行で分解し, 各文を Doc2Vec に入力することで文ベクトルを得る. 得られた文ベクトルの平均ベクトルを $\boldsymbol{v}_{\mathrm{goal}}$ とする. 基準となる文章を s とする.
2. 世代数を g, 最終世代を $g_{\max}$, 事前に生成した変換規則の集合を $R = \{R_i | i = 1, 2, \cdots, |R|\}$ とする. $g = 1$ とする.

表 1: 変換規則の例

変換元			変換先
直前の文節の末尾情報	文節の末尾情報	文節の先頭の品詞情報	表層表現
を-助詞-格助詞-一般	た-助動詞-*-*	動詞-自立-*	つかされた
は-助詞-係助詞-*	。-記号-句点-*	形容詞-自立-*	ないのだ。
文頭	の-助詞-連体化-*	名詞-一般-*	大部屋の
の-助詞-連体化-*	と-助詞-格助詞-引用	-名詞-数-*	一人と

3. s に対して構文解析し, 得られた文節集合を P^s とする. また, s を Doc2Vec に入力することで文ベクトル $\boldsymbol{v}_s$ を得る.
4. 変換規則集合 R から無作為に 1 つ変換規則を選択し, R_i とする.
5. P^s 中の j 番目の文節を $p_j^s (j = 1, 2, \cdots, |P^s|)$ とし, $j = 0$ とする.
6. j を $j+1$ と更新する. その後, $j > |P^s|$ ならば 4 に戻る.
7. 文節 p_j^s に対して, 3.1 節の 5 ステップと同様の手法で e_{j-1}^s, e_j^s を取得する.
8. R_i における適用条件の 2 要素が e_{j-1}^s および e_j^s と, R_i における変換元条件の要素が f_j^s とそれぞれ等しい場合, p_j^s を R_i における変換先の要素と置換する. いずれかの条件が 1 つでも等しくなければ, 6 に戻る.
9. P^s を結合することで新たな文章 s' を生成する. また, s' を Doc2Vec に入力することで文ベクトル $\boldsymbol{v}_{s'}$ を得る.
10. $\boldsymbol{v}_{\text{goal}}$ と $\boldsymbol{v}_s$ および $\boldsymbol{v}_{s'}$ のコサイン類似度を $\mathrm{S}(\boldsymbol{v}_{\text{goal}}, \boldsymbol{v}_s)$ および $\mathrm{S}(\boldsymbol{v}_{\text{goal}}, \boldsymbol{v}_{s'})$ とする.
11. $\mathrm{S}(\boldsymbol{v}_{\text{goal}}, \boldsymbol{v}_s) < \mathrm{S}(\boldsymbol{v}_{\text{goal}}, \boldsymbol{v}_{s'})$ であれば, s を s' に更新し, エリート文とする. そうでなければ, 確率 ϵ で s を s' に更新する. これは, 複数の箇所を変換することで, より類似度が大きい文章が生成できる可能性があるためである.
12. g を $g+1$ と更新する. $g < g_{\max}$ ならば, 3 に戻る. そうでなければ, エリート文を生成結果として文生成を終了する.

表 2: 実験条件

Doc2Vec に関するパラメータ	
分散表現の次元数	200
文脈窓	3
文生成に関するパラメータ	
ジャンル	恋愛-異世界, ファンタジー-現実世界
最大世代数 $g_{\max}$	500
確率 ϵ	0.1

4 実験

提案手法の有効性を確認するため, 提案手法により, 同一の文章からいくつかの文章を自動生成した. 生成された文に対して主観評価をした.

4.1 実験条件

表 2 に実験条件を示す. Doc2Vec に用いる学習データとして, 小説投稿サイトである「小説を読もう!」の年間ランキングから, 11 ジャンルの各上位 20 作品, 合計 220 作品を利用した. 変換規則の生成には各ジャンルの 年間ランキングが 1 位の作品を利用し, 同一のジャンルに変換するときのみ, その作品から生成された変換規則を用いることとした. また, 目標とするジャンルの文章には変換規則の生成に使用した作品を用いた. 各ジャンルで 1 作品のみを目標とすることで, その作品風な文章が生成されることが期待できる.

実験で用いる変換前の文章を以下に示す. なお, 変換前の文章は Doc2Vec に用いた学習データに含まれる作品 [6] から 1 文を抜き出したものである.

変換前の文章

懐かしい友達の面々を思い出しかけたところで、鼻先にポツン、と冷たいものを感じた。

表 3: 実験結果

ジャンル	生成文	コサイン類似度
恋愛 (現実世界)	懐かしい友達の電源を吐いたところで、海にポツン、と冷たいものを持った。	0.2121
	懐かしい友達の悪事を迫ったところで、海坊主にポツン、と冷たいものをかけた。	0.1742
ファンタジー (異世界)	懐かしい友達の DP を負ったところで、人間たちにポツン、と冷たいものを呪った。	0.6707
	懐かしい友達の DP を賭けたところで、粉々にポツン、と冷たいものを壊した。	0.6699

4.2 実験結果

表 3 に生成されたエリート文および実験途中での生成例を示す. 各ジャンルの上段がエリート文であり, 下段が実験途中での生成例である.

表 3 を見ると, エリート文はどれも意味的に妥当性がある文章とは言えない. これは文法的な条件で変換規則を適用できるかを決定してしまっており, 意味的な妥当性を考慮していないためであると考えられる. 意味的に妥当性のない文章は分散表現での類似度が減少することを期待していたが, 文章の分散表現での類似度以外にも, 単語の類似度を考慮する必要がある.

ファンタジー (異世界) のジャンルに注目すると, 目標とする作品とのコサイン類似度が他のジャンルと比較して大きな値を示した. これは「DP」という目標とした作品 [7] に固有な単語が文中に含まれたためだと考えられる. 目標となる作品の平均ベクトルとの類似度を考慮することで, 目標と同様の特徴を有した文章を生成できる可能性が示唆された.

5 まとめと今後の課題

本研究ではジャンルに依らないストーリーに基づいて, 任意のジャンルの小説を自動生成することを最終目標とし, 文章の分散表現と変換規則を用いたジャンル風文章の自動生成手法を提案した. 提案手法により生成された文章を主観評価することにより, 特定のジャンル特徴を有した文章を生成できる可能性があることを示した.

今後の課題として, Doc2Vec の学習データをより増加させることや, レビューや新聞記事のような小説以外のテキストデータを学習データとすることが挙げられる. 変換規則の適用条件に単語間の意味的類似度を考慮することや, 対象とする文自体の最適化手法の開発も今後の重要な課題である.

なお, 本研究は一部, 日本学術振興会科学研究補助金基盤研究 (C) (課題番号 26330282) の補助を得て行われたものである.

参考文献

[1] Kiyohito Fukuda, Saya Fujino, Naoki Mori, and Keinosuke Matsumoto. *Semi-automatic Picture Book Generation Based on Story Model and Agent-Based Simulation*, pp. 117–132. Springer International Publishing, Cham, 2017.

[2] 緒方健人, 佐藤理史, 駒谷和範. 模倣と置換に基づく超短編小説の自動生成. 第 28 回人工知能学会全国大会発表論文集, 2014.

[3] 宮崎千明, 平野徹, 東中竜一郎, 牧野俊朗, 松尾義博, 佐藤理史. 文節機能部の確率的書き換えによる言語表現のキャラクタ性変換. 人工知能学会論文誌, Vol. advpub, , 2016.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.

[5] gensim. `https://radimrehurek.com/gensim/models/doc2vec.html`.

[6] 筏田かつら. 静かの海 あいいろの夏、うそつきの秋. 宝島社出版, 2016.

[7] 月夜涙. 魔王様の街づくり！〜最強のダンジョンは近代都市〜. `http://ncodes.syosetu.com/n7637dj/`.