文テンプレートと対話型進化型計算を用いた文章の自動生成手法の提案

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.cs, mori@cs, matsu@cs}.osakafu-u.ac.jp

1　はじめに

2012 年 9 月, はこだて未来大学が星新一のショートショートをコンピュータで解析し, 新たなショートショートを生成するプロジェクト「きまぐれ人工知能プロジェクト 作家ですのよ」を開始すると発表した. このように近年, 計算機による小説の自動生成が人工知能や自然言語処理の分野で研究されている. しかしながら, 小説の自動生成には大きく分けて各文の妥当性と物語全体の一貫性という 2 つの問題点があり, 研究がそれほど進んでいないのが現状である. 文の妥当性に関しては, 人手で作成された文を利用することで, 妥当性があり意味の通じる文が生成できるようになりつつあり, 筆者らも限定状況下における解説文生成システム「なめ工房」を提案してきた. しかしながら, このシステムでは事前に人手で作成した文テンプレートを使用するため, 多様な文を生成するために非常に大きなコストがかかってしまう. また, なめ工房は 1 文のみを生成するシステムであり, 小説のように複数の文から構成された文章を生成できない.

以上の点を背景に, 本研究ではウェブ上の小説投稿サイトから取得した文章を基に作成した文章テンプレートを利用して文章を自動生成する手法を提案する. また, それらに進化型計算 (Evolutionary Computation, EC) の一種である対話型進化型計算 (Interactive Evolutionary Computation) を適用することで少数の文から構成される文章の生成システムを提案する. また提案手法を用いて生成された文に対して, 文の妥当性, 文章中の文の多様性, 文章としての整合性という 3 つの観点から評価することで, 提案手法について有効性を示す.

以下に本研究の構成を示す. 第 2 章で従来研究について述べ, 第 3 章で提案手法に用いた各要素技術について説明する. 第 4 章で提案手法について説明し, 第 5 章で評価実験について示す. 最後に第 6 章でまとめと今後の課題について述べる.

2　従来研究

筆者らはこれまでにユーザの嗜好を考慮した限定状況下における解説文生成システム「なめ工房」を提案してきた. このシステムには解説文に対話型進化型計算を適用することで, ユーザの嗜好を考慮した解説文の生成が可能であるという特徴がある. しかしながら, 解説文に特化した文テンプレート事前に人手で作成しているため, 作成できる解説文の種類を増やそうとすると文テンプレートを作成するコストが非常に高くなってしまうことが問題点として挙げられる. また, 「なめ工房」では解説文を 1 文と定義しているため, 複数の文から構成される文章が生成できなかった.

3　要素技術

ここでは, 本研究で用いる要素技術について詳述する.

3.1　形態素解析

形態素解析はテキストデータを形態素呼ばれる言語で意味を持つ最小の単位にまで分割し, 各形態素の品詞などを判定する自然言語処理における基礎的な技術である. オープンソースで利用できる日本語形態素解析器には京都大学黒橋・河原研究室で開発されている JUMAN や 非常に高速に解析できることが特徴の Mecab などが存在する. 形態素解析のアルゴリズムとしては, テキストデータをラティス構造で表現し, 各ノードとその接続についてコストを設定しておき, そのコストが最小となる経路を動的計画法で求めるビタビアルゴリズムが一般となっている.

3.2 熱力学的遺伝アルゴリズム

熱力学的遺伝アルゴリズム (Thermodynamical Genetic Algorithm, TDGA) は EC の一種であり, 明示的な個体群の多様性の制御を可能としたアルゴリズムである.

3.2.1 自由エネルギー最小化原理

温度 T で熱平衡状態にあるシステムでは, 状態の定常分布は自由エネルギー

$$F = \langle E \rangle - HT \tag{1}$$

を最小にする分布になることが知られており, これを自由エネルギー最小化原理と呼ぶ. ここで, $\langle E \rangle$ はシステムの平均エネルギー, H はエントロピーである. GA の観点からは, (1) 式の右辺第一項はシステムがエネルギー最小化 (GA における適応度の最大化) という目的を追求する項, 第二項はシステムの状態の多様性を維持する項と解釈でき, これらを温度 T をパラメータとして調和させたものと考えられる. TDGA は, 自由エネルギー F を最小化するように各世代で個体群を選択することによって, 明示的な多様性の制御を可能とした遺伝アルゴリズムである.

3.2.2 熱力学的遺伝アルゴリズムの概要

従来の GA で用いられていたルーレット式選択 [?] では, 適応度のみに注目して選択が適用されるため, 初期収束など多様性の制御が困難なために生じる問題があった. TDGA ではこの問題を解決するために, 個体群の多様性をエントロピーとして明示的に評価し, 個体群をその自由エネルギーが最小化となるように選択するルールを提案した. この選択は熱力学的選択ルールと呼ばれる [?]. 従来の選択ルールと同様に低いエネルギー E (高い適応度) を持つ個体は, (1) 式の自由エネルギーの右辺第一項の効果によって生き残る可能性が高くなる. 一方, 個体群において希少な遺伝子を多く持つ個体は, エントロピー H を高くするので, (1) 式の右辺第二項に寄与することにより自由エネルギーを減少させるため, やはり生存に有利な個体となる.

TDGA において, 温度 T は選択において多様性の維持にどの程度重点を置くかを調整するパラメータである. すなわち低温ではエネルギー値重視の, 高温では多様性重視の探索がされ, そのバランスを温度 T により明示的に制御することができるという利点がある.

4 提案手法

本研究では, 文テンプレートの作成コストを削減するために, 自動生成した文テンプレートと単語の上位概念を利用した文の自動生成法を提案する.

4.1 概要

高橋らは, 計算機による対話文の自動生成手法をチャットの対話ログから適切な応答文を抽出するログ型, あらかじめテンプレートとして用意された応答パターンに適宜単語を代入することで応答文を生成するテンプレート型, n-gram モデルを用いて応答文を生成する n-gram 型の 3 種類に分類している. 本研究では, この分類を対話文以外にも適用し, あらかじめ用意した文に適宜単語を代入, 変換, 挿入することで文を生成する手法をテンプレート型と定義し, 用意した文を文テンプレートと定義する. テンプレート型の特徴として, 意味の通じる文が作成されやすいことが挙げられる. しかしながら文テンプレートを人手で生成するには大きなコストがかかってしまう.

本研究ではこの問題点を解決するため, 既存の文に対して上位概念を利用した処理を施し作成した文テンプレートを用いて文を自動生成する. また, その文に対して IEC を適用することで文と文との関係性を考慮した文章を自動生成する手法を提案する.

4.2 文テンプレートの生成

本研究では文テンプレートに小説投稿サイト「小説を読もう！」から取得した文に前処理をしたものを利用する. 以下に前処理の流れを示す.

1. 小説は通常, 連続した文章から構成されているため, 句点を区切りとして 1 文に分割する.
2. 1 で取得した文を形態素解析にかけ形態素列を得る. 形態素解析には, JUMAN を利用した.
3. 形態素列の中から品詞が「形式名詞」, 「副詞的名詞」以外の名詞を抽出する.
4. JUMAN を用いて抽出された形態素のカテゴリを取得する. 表 1 にカテゴリの分類を示す. これは, JUMAN で得られるカテゴリ分類に姓名と地名を加えたものである.
5. 取得したカテゴリで対応する形態素を置換することで文テンプレートを生成する

表 1: カテゴリ分類一覧

カテゴリ名	例	カテゴリ名	例
人	学生, …	場所-施設	ビル, …
組織・団体	政府, …	場所-施設部位	天井, …
動物	犬, …	場所-自然	山, …
植物	桜, …	場所-機能	上, …
動物-部位	手, …	場所-その他	都市, …
植物-部位	葉, …	抽象物	思考, …
人工物-食べ物	パン, …	形・模様	円, …
人工物-衣類	ズボン, …	色	赤, …
人工物-乗り物	自転車, …	数量	複数, …
人工物-金銭	給料, …	時間	今日, …
人工物-その他	鉛筆, …	姓名	鈴木, …
自然物	石, …	地名	東京, …

6. 置換した形態素の数毎に文テンプレートを保存しておく.

7. 2 〜 6 の操作を取得した文のうち, 名詞の数 N_w が $1 \leq N_w \leq 5$ を満たす文に適用する. これは, 文を生成する際に代入する部分が多すぎると, 意味が通じない文が生成される確率が大きくなると思われためである.

4.3 文の生成

本研究では, ユーザの入力に応じて, 自動生成された大量の文テンプレートから文を自動生成する. 以下に文を自動生成するアルゴリズムを示す.

1. ユーザからの入力を得る. ユーザからの入力は単語のみとし, 1 〜 5 単語まで入力を許容している.

2. 入力に対して JUMAN を用いて各単語のカテゴリを得る. このカテゴリは, 文テンプレートを生成した際に利用したカテゴリ分類と同様のものを利用する.

3. 入力単語数と等しい数だけ置換された文テンプレート集合のなかから, 入力とまったく同じカテゴリを持つものを抽出し, 候補文テンプレート集合 $\mathbf{T}_c$ とする.

4. $\mathbf{T}_c$ に対してカテゴリと入力単語を置換することで妥当性が高いと考えられる文を生成する.