The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

文の分散表現に基づく小説の自動分割とストーリー展開の解析

Novel Segmentation Method based on the Distributed Representation of Sentences and Analysis Method of Story Developments

福田 清人 *1　森 直樹 *1　岡田 真 *1
Kiyohito Fukuda　Naoki Mori　Makoto Okada

*1大阪府立大学 工学研究科
Graduate School of Engineering, Osaka Prefecture University

Recently, the attempts to reproduce the mechanisms of human intellectual activities have attracted interest in artificial intelligence fields. The narrative creation is one of them. It is necessary for narrative creation and creative support by the computer to make it to understand human creations and their stories. However, there are few studies on story analysis by the computer. In this study, we propose the segmentation method of novels based on the distributed representation of sentences and the analysis method of story developments. As a result of computational experiments, we confirmed that the effectiveness of the proposed methods.

1. はじめに

近年, 人の知的活動の仕組みを計算機上で再現する試みが人工知能の分野で広く行われ, 大きな注目を集めている. 人の知的活動の 1 つに物語の創作がある. 人の感性に基づく創作物である物語はストーリーと表現媒体の 2 つの要素から構成され, それらの組合せによって小説や漫画などに枝分かれしていく. ここで, ストーリーは物語の内容であり, 表現媒体はその内容をどのような形で表現するかという表現方法である. 計算機による物語の自動生成の研究 [Pěrez 01][Ueno 14] は数多く報告されている. また, 近年では人と機械による共同作業に関する研究も注目を集めており, 人と計算機による物語の共同創作の研究 [上原 11] や創作支援の研究 [葛井 17] も報告されている. どちらの研究においても, 人の創作物を計算機に理解させることが非常に重要となる.

物語の自動生成や創作支援を実現するためには, 既存の物語を解析し, 人が物語を創作するうえで必要な知識や技術を計算機が理解可能な形で獲得する必要がある. 具体的には, 機械学習技術に基づく既存の物語に対する解析による有用な情報の抽出は必要不可欠な技術である. しかしながら, 物語の解析に関する研究は, 専門家の経験則に基づいて人手で情報を抽出する研究 [佐藤 10] が主流であり, 計算機による数値的な情報抽出や解析に関する研究はほとんど報告されていない.

なお, 本研究ではストーリーに焦点を当てる. 表現媒体が時代とともに姿を大きく変えることがある一方, 古くから存在するストーリーの典型的な構造が現代でもしばしば使用されることから, 表現媒体と比べてストーリーの方が時間経過に対してロバストなためである. また, 本研究における解析対象には小説を用いる.

以上を背景として, 本研究ではストーリーを計算機に理解させるための第一歩として, 文意を考慮した小説の自動分割手法およびストーリー展開の解析手法を提案する. 文意を考慮した文の分散表現に基づき, 小説文をストーリーが展開する部分で自動分割する. また, 自動分割された複数の小説文をシーンとみなし, 各シーンのベクトルを用いてストーリー展開が類似した部分を発見する.

連絡先: 福田清人，大阪府立大学 工学研究科，大阪府堺市中区学園町 1-1，072-254-9273，fukuda@ss.cs.osakafu-u.ac.jp

2. 関連研究

本研究と関連のあるいくつかの研究について説明する.

2.1 テキストセグメンテーションに関する研究

テキストデータをトピックなどの意味的なまとまりに分割するテキストセグメンテーションに関する代表的な手法に TextTiling[Hearst 97] がある. TextTiling はテキスト中のある 2 文間を基準として, その前後の文をあらかじめ設定した窓幅の分だけそれぞれ取得し, 得られた前後の文章に対して単語の出現頻度ベクトルの類似度を計算する. この操作を基準となる 2 文を動かしながら実行し, 得られた類似度の変化から文章境界を推定する手法である. TextTiling は文章内に出現する単語に基づいてセグメンテーションするため, 短い文章を対象とした場合には有効に機能しないことが知られている.

2.2 物語の解析に関する研究

物語の解析に関する研究では, 星新一の作品を構造分析の考えに基づき, テキストの時系列に着目して物語のパターン抽出をする研究が報告されている [佐藤 10]. しかしながら, 物語のパターンを抽出するためにはテキストを抽象化して分類する必要があるため, 人手によってしか解析できないという問題点が存在する.

3. 提案手法

本研究では, 物語の中でもストーリーという要素に着目して小説を解析する. ここで, ストーリーをイベントや登場人物の行動, 場所移動に伴う物語中の一連の状態遷移の時系列であると定義する. 小説をいくつかの文章の集合であると仮定すると, ある連続した 2 つの文章間の差が状態遷移であり, 冒頭から末尾までの連続した 2 文章間の差の分布がストーリーであるといえる. そこで, 小説中の文章を分散表現化して文章ベクトルを得た場合, 小説は文章ベクトルの時系列集合とみなすことができる. また連続する 2 つの文章ベクトルに何らかの演算子を適用した結果がその 2 文章間での差であり, 物語内の状態遷移を表しているといえる. そのため, 小説におけるストーリーは冒頭から末尾までの連続した 2 つの文章ベクトルにある演算子を適用した結果の時系列集合であると定義できる.

以上の観点から本研究では, 文の分散表現に基づく小説文の自動分割手法およびストーリー展開の解析手法を提案する. な

お, 文の分散表現の獲得には, これまでに提案してきた文の分散表現の獲得手法 [Fukuda] を改良した手法を用いる.

3.1 文の分散表現を用いた小説文の自動分割

TextTiling の考え方を基にして, 小説文の各文の分散表現に対して類似度を計算し, 類似度が極大となる 2 文を結合していく操作を, セグメント数が任意の数となるまで繰り返すことで小説を自動分割する手法を提案する. ここで 1 文単位での類似度計算をすると, 機械的な文分割により分割されてしまった不適切な文の前後を分割点と推測してしまう可能性がある. そこである 1 文に対して, その 1 文と前後窓幅分を含む文の分散表現の平均を類似度計算に用いるベクトルとするスムージング手法を導入する. 図 1 および図 2 に小説文のセグメンテーション手法の概要およびスムージング手法を示す. 以下に文の分散表現を用いた小説文のセグメンテーション手法のアルゴリズムを示す.

1. 獲得したいセグメント数を N_s, スムージング幅を N_w とする. ここで, 本節で用いるスムージング手法では基準となる文に対してその前後の N_w 文を含む $2N_\mathrm{w}+1$ 文をまとめてスムージングする.

2. 解析する小説を M 文の文集合とする.

3. 小説中の各文に対して, 文の分散表現の獲得手法により文の分散表現 $\boldsymbol{s}_i (i = 1, 2, \cdots, M)$ を獲得する.

4. 各セグメントに対応したセグメントベクトルを $\boldsymbol{d}_j$ $(j = N_\mathrm{w}+1, N_\mathrm{w}+2, \cdots, M-N_\mathrm{w})$, セグメントベクトルの集合を $\mathcal{D} = \{\boldsymbol{d}_{N_\mathrm{w}+1}, \boldsymbol{d}_{N_\mathrm{w}+2}, \cdots, \boldsymbol{d}_{M-N_\mathrm{w}}\}$ とする. また, 各セグメントに含まれる文数を b_j, この文数の集合を $\mathcal{B} = \{b_{N_\mathrm{w}+1}, b_{N_\mathrm{w}+2}, \cdots, b_{M-N_\mathrm{w}}\}$ とする. ここで,

$$\boldsymbol{d}_j = \frac{1}{2N_\mathrm{w}+1} \sum_{k=j-N_\mathrm{w}}^{j+N_\mathrm{w}} \boldsymbol{s}_k \tag{1}$$

$$b_j = \begin{cases} N_\mathrm{w}+1 & (j = N_\mathrm{w}+1, M-N_\mathrm{w}) \\ 1 & (\text{otherwise}) \end{cases} \tag{2}$$

である.

5. セグメントベクトル集合の連続した 2 つのセグメントベクトル $\boldsymbol{d}$ および $\boldsymbol{d}'$ の類似度 $f_\mathrm{sim}(\boldsymbol{d}, \boldsymbol{d}')$ を以下の式に従って計算する. ここで, γ は減衰率であり, セグメントが長文になりすぎないよう制御するための可調整パラメータである. γ は $0 < \gamma < 1$ を満たす実数である. また, b および b' はそれぞれセグメントベクトルに対応したセグメントに含まれる文数である.

$$f_\mathrm{sim}(\boldsymbol{d}, \boldsymbol{d}') = \gamma^{b+b'-2} \left(1 + \frac{\boldsymbol{d} \cdot \boldsymbol{d}'}{|\boldsymbol{d}||\boldsymbol{d}'|}\right) \tag{3}$$

6. 5 で求めた類似度が最大となった 2 つのセグメントベクトルを $\boldsymbol{d}_\mathrm{m}$ および $\boldsymbol{d}_{\mathrm{m}'}$ とし, それぞれに対応するセグメントに含まれる文数をそれぞれ $b_\mathrm{m}, b_{\mathrm{m}'}$ とする.

7. $\boldsymbol{d}_\mathrm{m}$ および $\boldsymbol{d}_{\mathrm{m}'}$ に対応するセグメントを結合し, 1 つのセグメントとする. その後, 以下の操作を適用することで各値を更新する. ここで, 記号 '$\rightarrow$' は 左式を右式で更新する操作を表す.

$$\boldsymbol{d}_m = \frac{1}{2N_\mathrm{w} + b_\mathrm{m} + b_{\mathrm{m}'}} \sum_{k=\mathrm{m}-\mathrm{N_w}}^{\mathrm{m}'+b_{\mathrm{m}'}+N_\mathrm{w}-1} \boldsymbol{s}_k \tag{4}$$

$$b_\mathrm{m} \rightarrow b_\mathrm{m} + b_{\mathrm{m}'} \tag{5}$$

$$\mathcal{B} \rightarrow \mathcal{B} \backslash \{b_{\mathrm{m}'}\} \tag{6}$$

$$\mathcal{D} \rightarrow \mathcal{D} \backslash \{\boldsymbol{d}_{\mathrm{m}'}\} \tag{7}$$

8. $|\mathcal{D}| > N_\mathrm{s}$ ならば, 5 に戻る.

9. $|\mathcal{D}| = N_\mathrm{s}$ の時, $\mathcal{D}$ および $\mathcal{D}$ の各要素に対応したセグメントを獲得する.

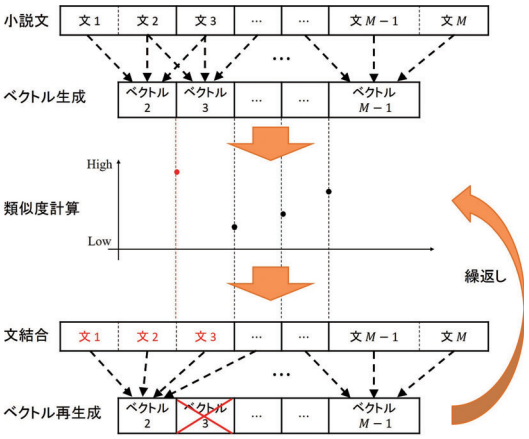

図 1: 小説文のセグメンテーション手法の概要

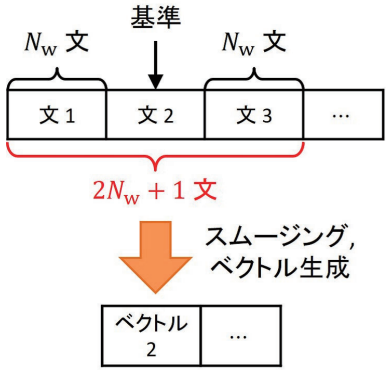

図 2: 小説文のスムージング手法の概要

3.2 文の分散表現を用いたストーリー展開の解析

3.1 節で説明した自動分割手法により得られたセグメントとそのセグメントに対応したセグメントベクトルを用いてストーリー展開を解析する. 本節では, 連続したセグメントに対応したベクトル間の差分を計算し, 得られた差分ベクトルを 1 つのストーリー展開とみなすことでストーリーを多次元数値空間上で表現する. 以下にストーリー展開の解析手順を示す.

1. 解析対象とする作品を複数用意する.

2. 3.1 節の自動分割手法により作品を自動分割し, 各作品のセグメントとそれに対応したセグメントベクトルを獲得する.

3. 各作品ごとに, 連続した 2 つのセグメントベクトルの差分を計算する.

表 1: 実験 1 における実験条件

$m_{\max}$	80
使用作品	太宰治 「走れメロス」 太宰治 「黄金風景」 芥川龍之介 「蜘蛛の糸」 芥川龍之介 「藪の中」 エドガー・アラン・ポー 「黒猫」

4. 得られた差分ベクトルに対して, 他作品から得られた差分ベクトルとのコサイン類似度を計算する.

5. 得られたコサイン類似度や差分ベクトルの各要素の割合に基づいて, 作品間でのストーリーの類似性や, 差分ベクトルとストーリー展開の関係性などを可視化しつつ解析する.

4. 実験

提案手法の有効性を確認するため, いくつかの実験をした. また, 提案手法を用いていくつかのストーリーを実際に解析することで, 得られる情報の特徴や傾向についての知見を得る. 実験 1 ～ 3 まで実施したが, 紙面の関係上, 実験 1 のみ示す. 実験 2 および実験 3 については発表時に述べる.

4.1 実験 1

文の意味的な類似性を考慮した文の分散表現を用いて, 小説のストーリー展開を可視化することができるかを確認する.

小説の各文に対して文の分散表現を獲得し, 次元圧縮手法である t-distributed Stochastic Neighbor Embedding (t-SNE) を用いて 2 次元空間に写像することで可視化する.

4.2 実験 1 の実験条件

表 1 に実験 1 の実験条件を示す. 実験 1 では青空文庫から取得した各作品を文単位に分割し, その中から $m_{\max}$ 単語以下の文を用いた.

4.3 実験 1 の結果と考察

図 3 ～ 図 7 に各作品の t-SNE による可視化の結果を示す. ここで, 各図において独立に t-SNE を適用しているため, それぞれの図の軸に関係性はないことに注意する. また, t-SNE はデータ間の距離を確率分布で表現することで次元を圧縮するため, 次元圧縮後の各軸には意味がないことにも注意を要を要する.

図 3 を見ると, 各文の分散表現が 2 つの分布パターンに属していることがわかる. これは, 「走れメロス」という作品は町での王様との会話シーンと, 村に戻ってまた町に戻るという移動シーンの 2 つに分けることができることから妥当であると考えられる.

図 4 および図 5 を見ると, すべての文の分散表現が次元圧縮後の 2 次元空間において, ある直線上に分布していることがわかる. このような分布をとる原因として, 「黄金風景」および「蜘蛛の糸」は場面転換時の風景描写などが淡白であることや, 作品全体を通して使用される文体や人物の口調なども一定であることが考えられる.

図 6 を見ると, 「藪の中」の各文の分散表現は先頭から末尾に向かって一定方向に展開していることがわかる. このことから, 「藪の中」は物語の後半部分で前半部分について回想することもなく, ストーリーが次々に展開していくと考えられる.

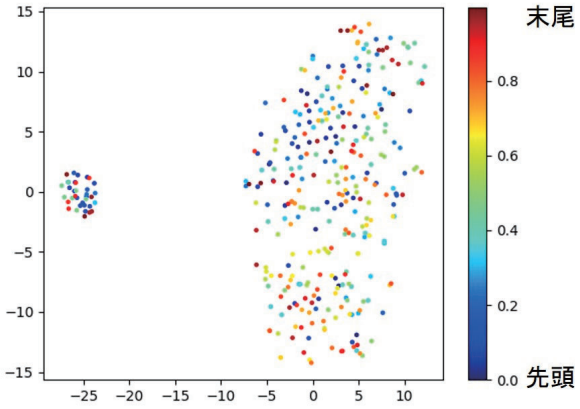

図 3: t-SNE による「走れメロス」の可視化結果

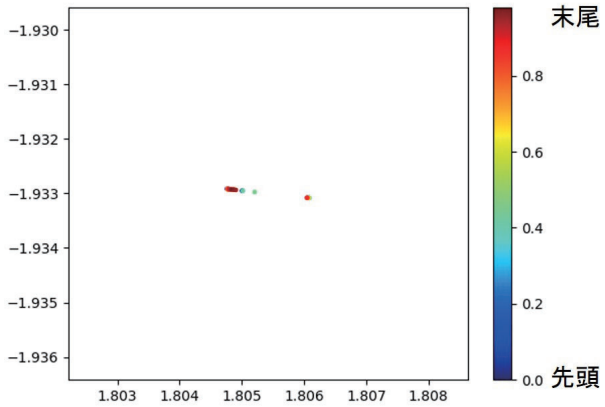

図 4: t-SNE による「黄金風景」の可視化結果

実際に, 「藪の中」という作品はある男が殺される事件に対して, 目撃者や容疑者, 被害者本人の霊などが事件について語るというストーリーであり, それぞれが自身の体験を語るだけで, 各人物が関わりあうシーンが存在しない.

図 7 を見ると, 文の分散表現が一様に分布していることがわかる. また, 「黒猫」という作品は主人公の身の回りで起きる事件と主人公の内面が交互に書かれている作品であり, 類似したストーリー展開を作品内で何度か繰り返す内容になっている. これらのことから, 似たストーリー展開を繰り返すことで, 時系列という観点から文の分散表現を可視化した結果, 一様に分布してしまっていると考えられる.

以上の結果から, 小説のストーリーやその展開に関して, 文の分散表現を用いた多次元数値空間上で作品の種類や特徴を解析可能であると考えられる.

5. むすび

本研究では, 文の意味的な類似性を考慮した文の分散表現に基づく小説の自動分割手法とストーリー展開の解析手法を提案した. 自動分割手法では TextTiling の考え方を基にして, 文の分散表現間の類似度が極大な部分を結合することで小説をシーンごとに自動分割した. ストーリー展開の解析手法では,

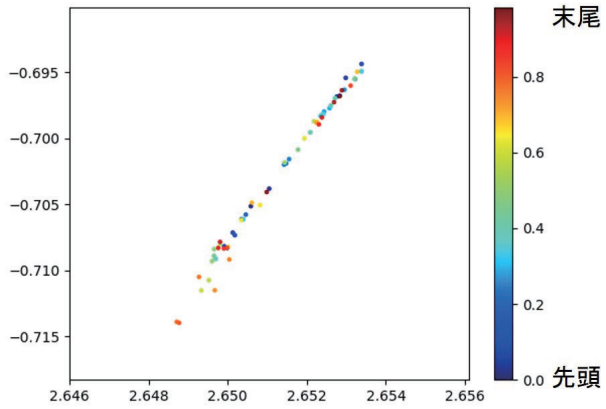

図 5: t-SNE による「蜘蛛の糸」の可視化結果

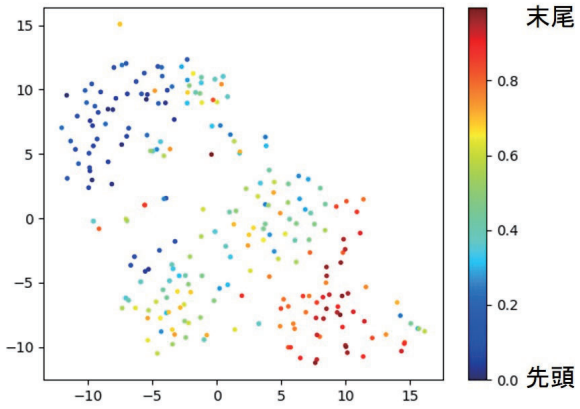

図 6: t-SNE による「藪の中」の可視化結果

自動分割手法で得られたセグメント間の差分をストーリー展開とみなし, 差分ベクトルの類似度からストーリー展開の類似性を解析した. 提案手法の有効性を確認するためのいくつかの実験により, 以下の知見が得られた.

- 小説文の分散表現を次元圧縮手法である t-SNE によって可視化することで, 小説のストーリーやその展開について多次元数値空間上で解析することができる.
- 提案手法を用いることで, 人手によるアノテートに頼ることなく小説文を意味を考慮したシーン単位に自動分割することができる.
- 提案手法により, ストーリー展開の類似性だけではなく, 文章構成の類似性も取得することができる.

今後の課題として提案手法の根幹となる文の分散表現の性能向上は最重要課題である. 階層的 LSTM や Attention 機構のような自然言語処理における有効性が示された技術を手法に導入することで更なる性能向上が期待される. また, 自動分割手法におけるスムージングの窓幅やシーンの分割数など解析対象の作品ごとに異なる可調整パラメータを最適化するための手法について検討する必要がある.

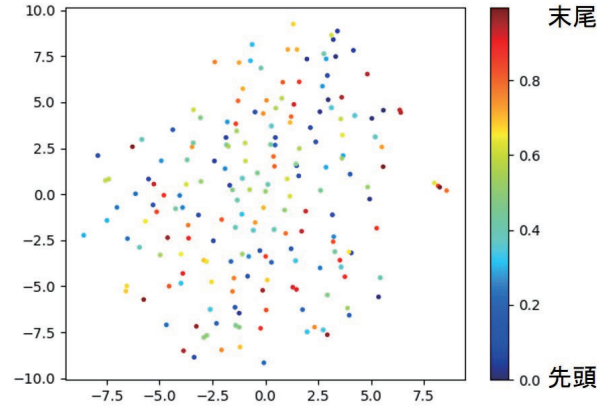

図 7: t-SNE による「黒猫」の可視化結果

謝辞

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

[Fukuda] Fukuda, K., Mori, N., and Matsumoto, K.: A Novel Sentence Vector Generation Method Based on Autoencoder and Bi-directional LSTM, in Prieta, de la F., Omatu, S., and Fernández-Caballero, A. eds., *Distributed Computing and Artificial Intelligence, 15th International Conference, DCAI 2018, Toledo, Spain, 20-22 June 2018*, Vol. 800 of *Advances in Intelligent Systems and Computing*, pp. 128–135

[Hearst 97] Hearst, M. A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Comput. Linguist.*, Vol. 23, No. 1, pp. 33–64 (1997)

[Pěrez 01] Pěrez, y R. P. and Sharples, M.: MEXICA: A computer model of a cognitive account of creative writing, *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 13, No. 2, pp. 119–139 (2001)

[Ueno 14] Ueno, M., Mori, N., and Matsumoto, K.: *2-Scene Comic Creating System Based on the Distribution of Picture State Transition*, pp. 459–467, Springer International Publishing, Cham (2014)

[葛井 17] 葛井 健文, 上野 未貴, 井佐原 均 : 質問集合とグラフに基づく物語全体の流れを管理可能な創作支援システムの提案, 第 31 回人工知能学会全国大会発表論文集 (2017)

[佐藤 10] 佐藤 知恵, 村井 源, 徃住 彰文 : 星新一ショートショート文学の物語パターン抽出, 情報知識学会誌, Vol. 20, No. 2, pp. 123–128 (2010)

[上原 11] 上原 大輝, 出水 ちあき, 宮里 洸司, 神里 志穂子, 野口 健太郎 : J-030 子どもの思考プロセス把握における物語自作システムの有効性検証 (HCS(2),J 分野:ヒューマンコミュニケーション&インタラクション), 情報科学技術フォーラム講演論文集, Vol. 10, No. 3, pp. 597–600 (2011)