Vol.2020-EC-55 No.21
2020/3/19

物語の感情曲線の抽出とその極値に着目したトピックの分析

番庄 智也[†1,a)]　藤井 叙人[†1,b)]　片寄 晴弘[†1,c)]

概要：これまでの物語論における研究では，個人が大量の本を読み，分類するといった研究手法が多く見られた．この手法は，個人の時間的，物理的限界があるといえる．一方で，コンピュータの処理能力の向上により，元々は人間が行なっていた処理をコンピュータが代替するというような例が見られるようになってきた．コンピュータが代替することで，時間的，物理的限界に縛られない研究が可能となる．本研究では Andrew Reagan らの感情曲線に関する研究について考察し，感情曲線を抽出し，その極値を収集する．収集されたテキストをトピック分析によって得られたベクトルによって階層型クラスタリングを行う．その結果，極値付近のテキストにおいては，キャラクターの背景・境遇が近い物語が属するクラスターがあることがわかった．

キーワード：物語論, 感情曲線, 自然言語処理

1. はじめに

古くから，物語の構造，役割について作家，哲学者の間で関心が持たれていた．従来の物語の研究では，個人が多くの書物を読み，読んだ書物に対して，分類をするということが一般的な研究手法だった．この研究手法は，個人ができるだけ多くの本を読む必要があるということから，(1) 主観的な分類になる，(2) 時間的，物理的限界が存在するといった問題が存在する．特に，年間 75，000 冊が出版され [1]，内容も多様化している現代において，個人が網羅的に多くの作品に触れるには限界がある．

一方で，情報処理技術の発展に伴い，コンピュータが人間の代替となって作業を行うという例が見られる．NTT docomo が開発する AI タクシー [2] は，顧客情報と位置情報を学習することで，タクシーの運転手に対してどこに顧客がいるかどうかを示すサービスである．これまでタクシーの運転手は経験と勘により，顧客がいると思われる場所に向かい，顧客を得ていたが，この AI タクシーでは，その経験と勘を機械学習で学習したコンピュータが代替となって務めることになる．

物語の研究についても同様に，コンピュータによって研究するという動きもある．Andrew Reagan らは，物語の感情曲線を取り出す研究を行った [3]．この研究は，物語の感情曲線は以下のような 6 つの基本的な型があるとしている．

- “Rags to riches” (上昇型)
- “Tragedy”，or “Riches to rags”(下降型)
- “Man in a hole”(下降，上昇型)
- “Icarus”(上昇，下降型)
- “Cinderella”(上昇，下降，上昇型)
- “Oedipus”(下降，上昇，下降型)

例えば，“Rags to riches”(上昇型) では，感情曲線が初めは低い幸福度のところにあり，上昇していくような物語である．また，Reagan らの研究は，感情曲線の基本的な型を発見したのみでなく，作家が幸福なイベントのときには幸福度の高い単語を，不幸なイベントのときには不幸度の低い単語を使う傾向にあるということを示唆している．

以上のことから，感情曲線の極値付近は，その物語の中で最も幸福または不幸な場面であることが考えられる．また，これらの幸福または不幸な場面を集積し，分類することで，共通するトピックや連想するキーワードにより，感情曲線をラベリングすることができると考えられる．本研究では，Reagan らの感情曲線抽出手法について基礎的な検証を実施する．その上で，感情曲線の値（極値付近）に着目してトピック分析およびクラスタリングを実施し，感情曲線自動抽出に基づく研究の可能性について議論する．

2. 物語と構造

物語に関する研究は，アリストテレスの“詩学”，ホラ

†1 現在，関西学院大学理工学部
Presently with Kwansei Gakuin University
a) fil26218@kwansei.ac.jp
b) fuji.nob195@gmail.com
c) katayose@kwansei.ac.jp

ⓒ 2020 Information Processing Society of Japan

ティウスの“詩論”まで遡ることができる．彼らは，当時の物語の表現形態である“演劇”を分析し，物語の機能や構造について著している．特に，アリストテレスは“詩学”において，悲劇的な場面で“神”が物語の文脈に関わらず，突然のように現れて事態を解決してしまうという当時の典型的なパターンを“機械仕掛けの神”(Deus ex Machina)と称して，批判した [4].

フォースターは，自著“小説と諸相”において，物語を構成する主要な要素は“ストーリー”，“登場人物”，“プロット”，“幻想”，“予言”，そして“パターンとリズム”の 7 つであると主張している [5]. また，フォースターは，“時間の進行に従って事件や出来事を語ったもの”を“ストーリー”と呼び [6]，“ストーリー”を因果関係を含めて考慮したものを“プロット”と呼んでいる [7].

Christpher Booker による“The Seven Basic Plots”では，今日に至るまでの演劇，小説などを分析し，物語を以下のような典型的な 7 つのストーリーに分類した[*1] [8]. 例えば，Overcoming the Monster には，登場人物に対して何らかの脅威となるような存在が現れ，登場人物たちが 1 度は敗北し，何かを奪われ，最後にはその脅威に打ち克ち，奪われたものを取り返す，というストーリーになる．

- Overcoming the Monster(例: “ジャックと豆の木”)
- Rags to Riches(例: “シンデレラ”)
- The Quest(例: “オディセウス”)
- Voyage and Return(例: “ガリバー旅行記”)
- Comedy(例: “真夏の夜の夢”)
- Tragedy(例: “アンナ・カレーニナ”)
- Rebirth(例: “クリスマス・キャロル”)

このように，古くから物語の構造に関する研究が行われていることがわかる．これらは，それぞれの著者が多くの書物を読み，それぞれの価値観やルールに基づいて分類している．特にアリストテレスの“詩論”や，ホラティウスの“詩学”は後の演劇にも影響を与えていると言われ，物語の構造に対する研究が実際に，物語の制作現場でも有効に働いている例もある．その具体例として，“Dramatica”を挙げることができる [9]. “Dramatica”は，ユーザが，今から作ろうとしている物語に関して，Dramatica から投げかけられる質問に答えることで，“Dramatica”から物語に関する知識が与えられる脚本支援ソフトである．この情報は，キャラクターの構成から物語の構造に関することまで多岐にわたる．また，“Dramatica”は実際にハリウッドの映画の脚本にも使われている．以上のことから，物語の構造についての研究は，現代の物語の制作現場でも生かされていると言える．

[*1] Christpher Booker は“plot”と表記しているが，それらは因果関係に重点が置かれているわけではないので，フォースターに倣い，この論文では”ストーリー”と表現する．

3. 実験手法

3.1 先行研究

Dodds らは，テキストの幸福度を以下の式で計算できるとした [10].

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i) p_i$$

ここで，N は文中に含まれる固有の単語群の数，i 番目の単語 w_i とその単語の文中で登場する頻度を表す f_i，その単語が持つ幸福度を $h_{avg}(w_i)$ とする．p_i は文中での i 番目の単語が現れる確率であり $p_i = f_i / \sum_{j=0}^{N} f_j$ で計算することができる．これは，テキストの幸福度をテキスト中に現れる単語の期待値としている．Reagan らは，この研究をもとに，小説の感情曲線を取り出す手法を提案した．図 1 は，感情曲線の抽出する手法である．

Reagan らの研究および本研究では，大量の小説のテキストデータが必要である．Reagan らは，英語の小説が大量に取得できるソースとして，Project Gutenberg[13] を使用した．Project Gutenberg では，著作権の切れた小説が有志によってテキストデータ化され，公開されている．本研究でも，Reagan らと同様，Project Gutenberg のテキストデータを活用する．Reagan らの研究では，英語の小説を，10,000 単語以上という条件で収集していたが，本研究では，12,000 単語以上とした．そのため，最終的に集まった書籍は，1,164 である．Reagan らの研究と同様，図 1 中のパラメータ，$N_w = 10,000$，$n = 200$ とした．なお，この研究では，1164 の書籍のうち，15 冊をテストデータとし，その他 1149 冊を LDA の学習用データとする．

3.2 感情曲線の抽出

3.2.1 感情曲線の抽出手法に関する検討

Reagan らの研究は，(1) コンピュータによる物語の研究の可能性，また (2) 作家による恣意的な感動のデザインの可能性を示唆したことで非常に先駆的であり，工学的な物語の研究をより進めたと言える．特に，(2) 作家による恣意的な感動のデザインの可能性を示唆した点は物語の研究において寄与する部分が大きい．この研究の示唆するところは，幸福な場面では幸福度の高い単語が，不幸な場面では幸福度の低い単語が意図的に選ばれているということである．

一方で，Reagan らの研究では，Dodds らの手法でも示されている通り，テキスト中に含まれる単語の幸福度の期待値をテキストの幸福度としているため，文法的による意味の変化を考慮していないことが課題として挙げられる．日常会話や物語においては，例えば，“not”や“few”といった数量詞を指定することで，意味が反対になるなど期待値だけでは評価できないような例も存在する．そこで，文法を

ⓒ 2020 Information Processing Society of Japan

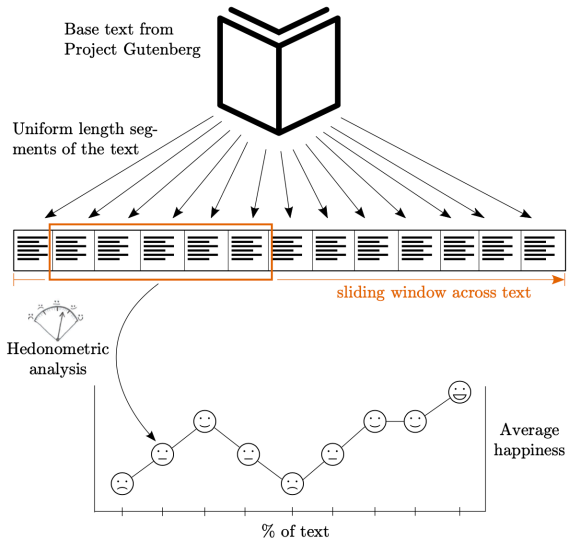


図 1: Reagan らによる物語から感情曲線を抽出する手法
単語数 N のテキストデータ化された小説を N_s のセグメント (segment) に分ける．N_w は図のオレンジに囲われた部分の単語数である．この時，n をスライドする回数とすると，$N_s = (N - (N_w + 1))/n$ を満たす．この時， セグメントに従ってウィンドウをスライドさせることで，ウィンドウごとの幸福度を取り出すことで，感情曲線を取り出すことができる．

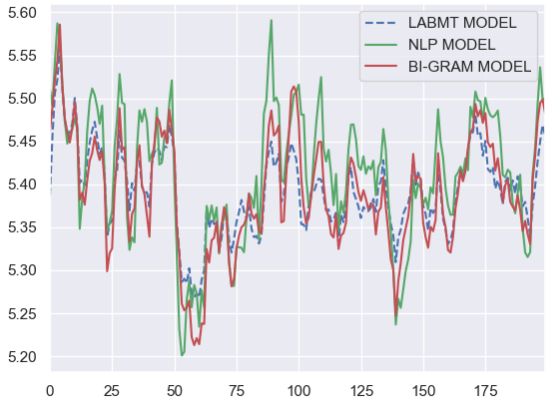


図 2: 物語の感情曲線を 3 つのモデルで比較した例

考慮したモデル (NLP Model)，Reagan らが使用したモデル (LABMT Model)，そして比較のため，Louis-Philippe Morency らが映像中の単語について検証したときに使われた [12] モデルをもとにした機能語に近接する単語の幸福度に機能語固有の値を積算して算出するモデル (BIGRAM Model) の 3 つで比較を行う．

図 2 は NLP Model，LABMT Model，および BIGRAM Model の 3 つによって抽出した感情曲線を描画した結果である．全体的に，NLP Model は LABMT Model から考えると大きな振れ幅になっており，BIGRAM Model はほとんど NLP Model と LABMT Model の中間または LABMT Model とほとんど同じ値を行き来している．また，感情曲線の上下の変化量については互いにほとんど同じ (上昇，下降がほとんど同じ) である．Reagan らの示した感情曲線はテキストの幸福度は相対的な値であり，絶対的な幸福度との相関が示されてはいない．そのため，ここで感情曲線において重要なのは値の変化量であると思われるため，変化量に着目したとき，互いに大きな変化があるとは言えない．そこで，本研究では，感情曲線の抽出手法を，最も計算が平易な Reagan らによる手法を用いることにする．

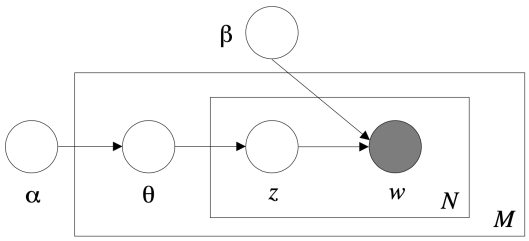


図 3: LDA のグラフィカルモデル [14]
α，β はハイパーパラメータ．N は単語集合．$z \longrightarrow w$ は，各トピック内の単語への積を表す．また，θ は，トピックが選ばれる確率を表す．M はトピック数．文書 D は，このモデルに従って生成されると想定し，実際に用いられる場合では，トピック数を指定し，z を変化させることで，w の重みを推定する．

3.3 トピック分析

3.3.1 Latent Dirichlet Allocation によるトピックモデルの作成

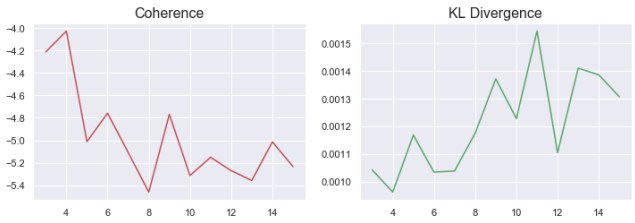
トピック分析では，LDA を用いる．図 3 は LDA のグラフィカルモデルである．

感情曲線のラベリングを行うため，ラベルとなるトピックを出力する LDA のモデルを作成する．トピック抽出モデルを作成するにあたって，同じ文書の中で現れる回数 2 回以上，また，全体の書籍の中で登場する割合が 75%未満となるように，その他の単語は取り除く．

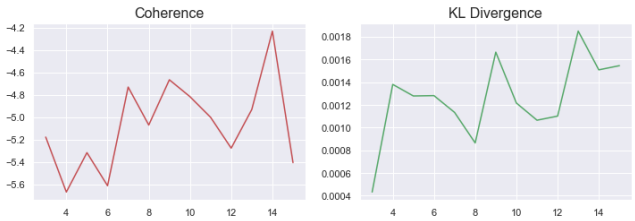
3.3.2 トピック数の決定

現在，感情曲線から取り出すことができるトピックが，いくつ程度あるか，わかっていない．そのためトピック数の検証を行う必要がある．そこで，トピック数ごとにモデルを生成し，生成されたモデルの中で最も有効なモデルを選ぶことにする．LDA のモデルの有効性を測る指標として，Coherence，KL Divergence による単語群の確率分布の類似度によるものがある [15].

図 4 は各評価指標による値とトピック数の関係である．トピック数は 3 から 15 へ変動する．また各評価指標による値を計測するために，予め学習したモデルに新たにテス



(a) 感情曲線の最小値の Coherence，KL Divergence による結果



(b) 感情曲線の最大値の Coherence，KL Divergence による結果

図 4: それぞれの極値のトピックモデルによる評価指標の結果

トのために残していたテストコーパスによって計測した．

まず，Coherence について述べる．Coherence はトピック中における単語が，人間にとって解釈しやすいかどうかを示す．Coherence は値が高いほど，人間にとって理解しやすい単語群であることを示す．Coherence のグラフを見ると，トピックが 4 の時最大となり，それ以降，ずっと低い値を取っている．

次に，KL Divergence のよる評価について述べる．KL Divergence は 2 つの確率分布の差異を表すものである．差異が大きければ大きいほど，KL Divergence の値は大きくなる．2 つの確率分布 $P(x)$，$Q(x)$ が与えられた時，KL Divergence は次のように計算される．ここで x は離散的とする．

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

LDA は各トピック中に含まれる単語全てを確率分布としてそれぞれ取り出すことができる．そこで重複しない 2 つのトピックのペア全てについて，それぞれペアごとに KL Divergence をとり，それらの平均をそのトピックの KL Divergence の値とする．このようにして，全てのトピック数について計測しプロットしている．佃らは Mr. Children の歌詞におけるトピック数を決定する上で，KL Divergence が最も低い値になる時のトピック数を Mr. Children のトピック数としている [15] ことから，本研究でも同様にその評価指標として KL Divergence も考慮している．

図 4(a) で注目するべき点はトピック数 4 を指定したときである．図 4(a) において，トピック数が 4 において Coherence が最も高く，KL Divergence は最も小さい値を取っている．佃らの研究によれば，これは最も良いトピック数となっていると考えられる．

また，感情曲線の最大値のウィンドウによって構成したモデルの評価指標による結果を示したのが図 4(b) である．図 4(a) のように区別できるような点が存在しない．本研究では，KL Divergence がトピック数 3 から 4 にかけて急に上昇し，トピック数 4 から単調上昇に転じていることからトピック数 3 に着目する．

Topic 0	Topic 1	Topic 2	Topic 3
king	god	nature	love
god	woman	love	god
father	king	god	woman
love	illustration	father	king
yes	lord	mother	nature

表 1: 最小値ウィンドウによるトピック数 4 の時のそれぞれのトピックにおいて生起する確率の高い 5 つの単語

Topic 0	Topic 1	Topic 2
god	god	god
king	king	mother
woman	yes	king
lord	woman	son
mother	illustration	gold

表 2: 最大値ウィンドウによるトピック数 3 の時のそれぞれのトピックにおいて生起する確率の高い 5 つの単語について

3.3.3 それぞれのトピックで生起しやすい単語の可視化

表 1，2 はそれぞれ，最小値ウィンドウ，最大値ウィンドウによるトピックから生起する確率が高い単語を上位 5 単語を並べたものである．また，感情曲線の最小では，“god” と “love” が生起する確率がとても高くなっている．

3.3.4 クラスタリング

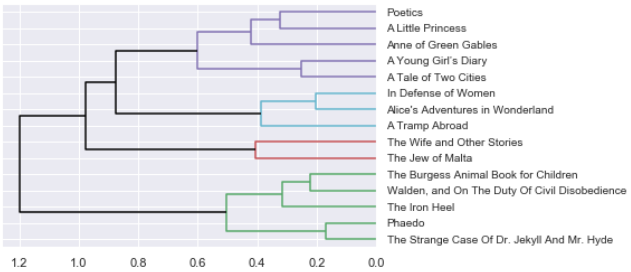
次はテスト用に用いる 15 の作品について，各トピックの値をベクトルとして，Ward 法を用いた階層型クラスタリングを行う．図 5 は，クラスタリングによる結果である．

興味深いのは，感情曲線の最大値によるクラスタリングでは，少女が主人公の物語 (“A Little Princess” や “Alice’s Adventures in Wonderland”) が同じクラスに属している．また，最小値によるクラスタリングでは，“A Little Princess” と “Anne of Green Gables” が同じクラスに属している．一方で，“A Little Princess” に最も近いものとして，アリストテレスの作品の “Poetics” があげられる．

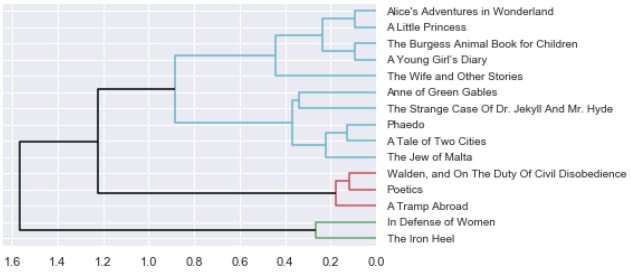
4. 考察

4.1 各トピックの分類に関する考察

3.3.3 節では，3.3.1，3.3.2 節で構成したモデルの各トピックモデルから生起する確率の高い単語について示した．この生起する確率の高い単語から，各トピックについてラベリングを行うという研究の例があるが [15]，表 1，2 から連想されるような単語はないと考えられるため，この段階でのラベリングは行わないことにする．



(a) 感情曲線の最小値のトピック数 4 の時のクラスタリング結果



(b) 感情曲線の最大値のトピック数 3 の時のクラスタリング結果

図 5: 感情曲線の極値におけるクラスタリング結果

興味深い点としてあげられるのは，どちらのモデルでも“god”，最小値のウィンドウによるモデルでは加えて“love”がどのトピックにも共通して現れるという点である．これは収集した Project Gutenberg データセット固有のものであると考えられる．特に聖書をモチーフにした物語に多いが，“god”と“love”は“God loves you”やトルストイの小説“Where Love Is，God Is”に代表されるように“god”と“love”は同時に使われることが多いと見られる．今回書籍を収集するソースとなった Project Gutenberg は，著作権切れの物語が集まっており，比較的歴史のある書物が多い．そのため，“god”や“love”のような比較的結びつきやすい単語が生起しやすくなっていると考えられる．このように生起しやすい単語を見ることで民族学的な研究の余地を見出すことができると考えられる．

4.2 クラスタリングによる考察

3.3.4 節では，3.3.1，3.3.2 節で構成したモデルによってテストコーパスをトピックで分類し，そのクラスタリング結果を示した．また，この節で示したとおり，感情曲線の最大値によるクラスタリングでは，少女が主人公の物語が同じクラスに属していることがわかる．また，最小値によるモデルでは，“A Little Princess”と“Anne of Green Gables”が同じクラスに属していることを示した．ここで，“A Little Princess”は，主人公サーラの最も不幸なイベントとして，父親との死別がある [16]．また，“Anne of Green Gables”は日本語訳では“赤毛のアン”と訳され，こちらも両親との死別が主人公アンの背景にある．このように，最小値によるモデルでは，主人公のキャラクターの背景が同じクラスに属している様子が見られる．

5. おわりに

本研究では，感情曲線自動抽出に基づく研究の可能性について議論するため，次の 2 つの視点について研究を行った．感情曲線自動抽出の手法に関する視点，そして自動抽出された感情曲線が持つ物語における意味に関する視点である．

まず，感情曲線自動抽出の手法に関する視点について述べる．Reagan らの研究によれば小説における感情曲線自動抽出の手法は，テキストに含まれる単語の幸福度の期待値を幸福度とする，という研究に基づくものだった．この手法は非常に平易な計算を可能としながら，他方で文法を考慮していないという課題を含む．物語においてキャラクターが“怒っている (angry)”か，“怒っていない (not angry)”かはその物語の幸福度を決める上で重要であると考えられるからである．そこで本研究では，Reagan らのモデルの他に，文法を考慮するモデルと文法は考慮しないが機能語を考慮するモデルの 2 つのモデルを仮定し，検証を行った．結果は 3.2.1 節でも示した通り，3 つのモデル間に大きな差を見出すことはできないと考えられる．一方で，感情曲線自動抽出の手法に関する重要な課題として，抽出された幸福度が絶対性を持たないということが挙げられる．この幸福度が絶対性を持たないということは，感情曲線における最大値，最小値を比較することは可能だが，例えば極大点または極小点同士で比較したとき，どちらの方が幸福かを示す指標にはならないということを示す．この問題はすなわち，感情曲線の起伏がそのまま登場人物の心の昂りや感情の上下を示さないということが考えられ，議論の余地があると言える．

次に，自動抽出された感情曲線が持つ物語における意味に関する視点について述べる．この視点は，感情曲線の各点がどのような意味を持つのかについての議論を深めることになろう．本研究では，感情曲線の極値に焦点を当て，その付近では何が起きているのかについて検証した．当初は，トピックモデルの生起する確率の高い単語に基づいて，連想する事柄によって分類することを試みたが，生起する確率の高い単語によって連想する事柄をトピックとする手法ではトピックのラベリングが可能とは言えなかった．しかし，各テキストをトピックモデルを通して得られたベクトルによってクラスタリングをすることで，読んだことのある物語から連想可能な事柄が存在するクラスタが現れた．以上のことから，クラスタリングし，クラスタに含まれる要素について連想され得る事柄によって物語を分類するという方向性があると考えられる．

以上 2 つは，研究する前段階で想定していた仮定に対する結果だが，本研究で面白いと思われた結果として，トピックモデルから生起する確率の高い単語から民族性を窺わせる結果が得られたことである．Reagan らは，“物語論

ⓒ 2020 Information Processing Society of Japan

(Narratology)”を研究することと“民俗学 (Folksology)”には関連があると主張している [17] が，本研究結果はその裏付けとなるような事実を窺わせるものだったと考えられる．

本研究結果は，感情曲線自動抽出の手法と自動抽出された感情曲線が物語において持つ意味を示す上で，様々な視点と考察を与えたと言える．今後は，これらの研究結果をもとに，感情曲線自動抽出の手法が持つ課題点の解決，感情曲線が持つ物語に対する他の意味，また自動抽出された感情曲線を足がかりとする物語論と民俗学との関連について議論を深めたい．

参考文献

[1] https://www.sokunousokudoku.net/media/?p=1690，2020 年 2 月 21 日確認
[2] https://www.nttdocomo.co.jp/biz/service/aitaxi/ , 2020 年 2 月 20 日確認
[3] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds:“The emotional arcs of stories are dominated by six basic shapes”(2016)
[4] アリストテレース (著)，ホラーティウス (著)，松本 仁助 (翻訳)，岡 道男 (翻訳):“アリストテレース詩学／ホラーティウス詩論”，岩波文庫 (1997)
[5] E.M. フォースター，中野康司 (訳):“E.M. フォースター著作集 8 小説の諸相”，みすず書房 (1994)
[6] 同上，p.40
[7] 同上，p.129
[8] Christpher Booker,“The Seven Basic Plots”, Continuum Intl Pub Group(2006)
[9] http://dramatica.com/ ，2020 年 2 月 20 日確認
[10] Peter Sheridan Dodds , Kameron Decker Harris，Isabel M. Kloumann，Catherine A. Bliss，and Christopher M. Danforth: “Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter”(2011)
[11] 同上 p.2
[12] Louis-Philippe Morency，Rada Mihalcea and Payal Doshi: “Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web”(2011)
[13] “Project Gutenberg”, https://www.gutenberg.org/, 2020 年 2 月 20 日確認
[14] David M. Blei，Andrew Y. Ng，and Michael I. Jordan: “Latent Dirichlet Allocation”(2003)
[15] 佃 洸摂，石田 啓介，後藤 真孝,“Lyric Jumper : アーティストごとの歌詞トピックの傾向に基づき様々な歌詞に出会える歌詞探索サービス”(2019)
[16] フランシス・ホジソン・バーネット，畔柳和代 (訳): “小公女”，新潮文庫 (2015)
[17] [3]，Reagan ら，前掲書，p.1

ⓒ 2020 Information Processing Society of Japan