ARG WI2 No.11, 2017
WI2-2017-19

# 小説の読書行動と逐次的なあらすじの特徴との関連の分析

森 晴菜[†, a]　山西 良典[††, b]　西原 陽子[††, b]　福本 淳一[††, b]

† 立命館大学大学院情報理工学研究科　†† 立命館大学情報理工学部

a) *h_mori@nlp.is.ritsumei.ac.jp*　b) {*ryama@media, nisihara@fc, fukumoto@media*}*.ritsumei.ac.jp*

**概要**　読書中断後，読書を再開する際に既読内容のあらすじを提示すれば，既読内容を短時間で振り返ることができ，本来の読書時間を削減することなくスムーズに続きを読み始められる．本稿では，読書再開時に必要とされるあらすじについて調査するため，人手によるあらすじ作成実験を行った．得られたあらすじについて，あらすじ文の小説本文中での出現位置や，読書進度に応じたあらすじの変化に着目し，その特徴を分析した．また，毎回の読書量や読書時間などの読書活動状況とあらすじとの関連についても考察した．

**キーワード**　あらすじ特徴，読書行動，逐次的あらすじ

## 1　はじめに

ドラマやアニメ，週刊連載の小説や漫画などでは，視聴者・読者に前回の内容を想起させる手段として，あらすじが用いられることが多い．長編小説の読書時には，読書中断後，読書を再開する際に物語の過去の内容を忘れてしまい，読み返しを行う場合がある．ここでも，あらすじは有効に働き，既読内容の短時間での振り返りや，スムーズな読書再開が期待される [1].

あらすじは，特定の単語（キャラクターやイベント）に関連した事柄を既読範囲から抽出して要約することで構成される．この観点から見れば，一般的な Query focused summarization（QFS）の一部として捉えることができる．一方で，読書の進行に応じて要約対象となる文章が更新されて，あらすじが変化する．この観点から見れば，Update summarization（US） [2] の一種としても捉えられる．つまり，読書進度に応じたあらすじ生成は，QFS と US の両性質を併せ持つと考えられる．あらすじ生成に関する研究としては，今野らの小説要約 [3] や野崎らの物語理解システム [4] がある．単語の重要度に関しては白鳥ら [5] や田島ら [6] の各ジャンルにおける重要単語判定や，前田ら [7] の小説テキストにおけるネタバレ単語についての調査がある．

本稿では，読書再開時に必要とされるあらすじについて調査するため，人手によるあらすじ作成実験を行う．得られたあらすじについて，あらすじ文の小説本文中での出現位置や，読書進度に応じたあらすじの変化に着目し，その特徴を分析する．

## 2　人手によるあらすじ作成実験

本研究で対象とする小説は，本文が 20,000 字以上の小説テキストとした．20,000 字以上の小説を読む際には読書の中断が複数回あると考えられ，あらすじの有用性が期待されるためである．文章量の多い小説は一般的に登場人物数も多いと想定され，あらすじによる情報整理の意義があることも理由として挙げられる．あらすじ作成では，青空文庫において公開中の小説「こころ」,「銀河鉄道の夜（新潮文庫版）」,「地獄変」のテキストデータを用いた．表 1 に，これら 3 作品の小説 ID と著者名をはじめとするメタ情報を示す．同表中の文字数，単語数，文数，段落数は，テキストデータから題名，著者，脚注，底本情報，見出しを削除した後，本文に含まれる数になる．ID $N_1$ については，文字数を他の 2 作品とおおよそ均等にするため,「上　先生と私」部分のみを抜粋して使用した．使用した 3 作品は，2016 年通年の XHTML 版青空文庫アクセスランキング（全 13,969 作品，2016.12.31 時点）において，それぞれ 2 位，11 位，58 位と高順位にある作品であり，各作品の著者についても著者毎の作品アクセス数合計の上位 3 名である．

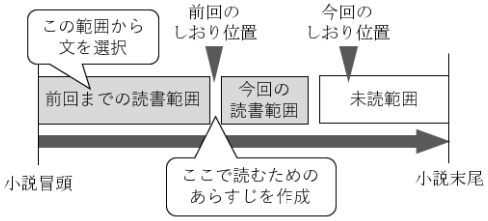


図 1 あらすじ文選択範囲

あらすじ作成実験は，割り当てられた各作品の内容を知らない 10 代後半から 20 代の男女合計 22 名により行われた．各作品を用いたあらすじ作成実験の被験者数を，表 1 中に示す．被験者には，実験期間中は小説のあらすじなどのネタバレとなりうる情報を得ないよう指示した．本実験は，以下の 3 手順の反復により行われた．

1. 小説を閲読させる
   読書開始時刻・読書終了時刻と，読書を中断してしおりを挟んだ位置を記録させる．

表 1 実験に使用した小説のメタデータと各小説を用いた実験の被験者数

| 小説 ID | 著者 | 題名 | 文字数 | 単語数 | 文数 | 段落数 | 被験者数 |
|---|---|---|---|---|---|---|---|
| $N_1$ | 夏目漱石 | こころ（上） | 49,103 | 32,415 | 1,779 | 282 | 7 |
| $N_2$ | 宮沢賢治 | 銀河鉄道の夜（新潮文庫版） | 38,212 | 22,963 | 1,100 | 158 | 8 |
| $N_3$ | 芥川竜之介 | 地獄変 | 26,103 | 18,313 | 480 | 98 | 7 |

表 2 各小説における被験者によるしおり挿入箇所

| | | 章節区切り | 空行 | 段落間 | 文間 | しおり総数 |
|---|---|---|---|---|---|---|
| しおりの数 | $N_1$ | 20 | 0 | 0 | 0 | 20 |
| | $N_2$ | 14 | 1 | 5 | 4 | 24 |
| | $N_3$ | 13 | 0 | 1 | 0 | 14 |

表 3 選択されたあらすじ文の位置

| 小説 ID | しおり直前の読書範囲のみから選択 | しおり直前の読書範囲と小説冒頭から選択 | 既読範囲から網羅的に選択 |
|---|---|---|---|
| $N_1$ | $S_2$, $S_3$, $S_4$, $S_6$, $S_7$ | $S_1$ | $S_5$ |
| $N_2$ | $S_9$, $S_{10}$, $S_{11}$, $S_{12}$, $S_{14}$, $S_{15}$ | $S_{13}$ | $S_8$ |
| $N_3$ | $S_{17}$, $S_{18}$, $S_{19}$, $S_{20}$ | $S_{16}$, $S_{21}$, $S_{22}$ | |

表 4 被験者ごとのしおり直前文のあらすじへの選択確率

| 小説 ID | $p_i = 0.00$ | $0.00 < p_i \leq 0.50$ | $0.50 < p_i \leq 1.00$ |
|---|---|---|---|
| $N_1$ | ($S_1 : 0.00$), ($S_3 : 0.00$), ($S_4 : 0.00$), ($S_5 : 0.00$), ($S_6 : 0.00$) | ($S_7 : 0.50$) | ($S_2 : 1.00$) |
| $N_2$ | ($S_{11} : 0.00$), ($S_{12} : 0.00$), ($S_{15} : 0.00$) | ($S_{10} : 0.50$), ($S_{13} : 0.33$) | ($S_8 : 0.75$), ($S_9 : 0.60$), ($S_{14} : 0.67$) |
| $N_3$ | ($S_{16} : 0.00$), ($S_{17} : 0.00$), ($S_{18} : 0.00$), ($S_{19} : 0.00$), ($S_{20} : 0.00$) | ($S_{21} : 0.50$) | ($S_{22} : 1.00$) |

2. 内容確認のための質問に回答させる
   質問は 1 作品につき 10 問用意しており，毎回の読書範囲において回答可能な質問を読書終了ごとに提示する．
3. あらすじを作成させる
   既読範囲から 8 文を選択し，小説本文中での出現順に並べたものをあらすじとして扱う．

あらすじに選択する文数は，作品名・著者名の異なるライトノベル 34 冊のあらすじ文数を調査した結果に得られた平均値 8.17647 を小数第一位で四捨五入した値から設定した．

あらすじ作成時には，前提として，被験者に以下の 2 点を伝えた．

- あらすじとは,「読者が続きを読む際に必要な情報」のことを指す．
- 作成するあらすじは，今回の読書範囲を読む直前に読むものであり，前回までの読書範囲に含まれる文を抜き出して作成する．

読書中断時におけるあらすじ文の選択範囲を，図 1 に示す．今回読書を中断した位置，すなわち今回のしおり位置，まで読んだ際には，小説冒頭から前回のしおり位置までの範囲から，あらすじとする文を選択する．つまり，2 回目読書後には小説冒頭から 1 回目のしおり位置までの範囲，3 回目読書後には小説冒頭から 2 回目のしおり位置までの範囲から，それぞれあらすじ文を選択する．1 回目読書終了時には前回のしおり位置が存在しないため，あらすじ作成は行わない．

内容確認のための質問に対して，被験者 22 名全員が 80%以上の正答率を示した．すべての被験者が小説の内容を理解した上であらすじを作成したと考え，被験者 22 名の作成したあらすじすべてを，有効な参考情報として扱った．

## 3 あらすじの考察

被験者によって作成されたあらすじと被験者の読書活動の結果，またそれらの関係について，以降で述べる．考察では,「しおり位置の傾向」「作成されたあらすじ文の抽出源」「読書間隔とあらすじ」「あらすじ中の意味情報」の観点を用意した．

### 3.1 しおり位置の傾向

被験者がしおりを挟んで読書を中断した位置を表 2 に示す．同表中では重複を認めておらず，章節区切り，空行，段落間，文間の順に判定を行い，該当項目が存在した時点で判定を終了した．3 作品すべてにおいて，しおりの半数以上が章節区切りとなる位置に挿入された．また，章節区切りと空行は段落間に包含されるため，どの作品においても，しおり総数の 83%以上は段落間に挿入されたことになる．

文章において，段落は 1 つの意味を持つ文のまとまりであり，段落ごとに主題を持つ．段落は章節における物語の区切りであると考えられ，読者が小説の読書を中断するタイミングは物語の区切り位置であると示唆された．

### 3.2 あらすじ文の抽出源

被験者の毎回の読書範囲と小説中のあらすじ文の位置との関係を分析した．選択されたあらすじ文の位置について，しおり直前の読書範囲のみから選択されたパターン，小説冒頭としおり直前の読書範囲から選択されたパターン，既読範囲全体から網羅的に選択されたパターンの 3 通りの傾向が見られた．小説冒頭とは，各被験者の

表 5 読書間隔とあらすじ文の位置の関係（3 回目読書前のあらすじ）

| 小説 ID | 1 日未満 | 1 日以上 3 日未満 | 3 日以上 7 日未満 | 7 日以上 14 日未満 | 14 日以上 30 日未満 | 30 日以上 |
|---|---|---|---|---|---|---|
| $N_1$ | | $(S_2 : 0.00)$, $(S_5 : \mathbf{0.63})$, $(S_6 : 0.00)$ | $(S_1 : \mathbf{0.38})$, $(S_4 : 0.00)$, $(S_7 : 0.00)$ | $(S_3 : 0.00)$ | | |
| $N_2$ | $(S_{10} : 0.00)$, $(S_{11} : 0.00)$, $(S_{13} : \mathbf{0.63})$, $(S_{15} : 0.00)$ | | | $(S_9 : 0.00)$ | $(S_8 : \mathbf{0.25})$ | $(S_{12} : 0.00)$, $(S_{14} : 0.00)$ |
| $N_3$ | $(S_{20} : 0.00)$ | $(S_{18} : 0.00)$, $(S_{19} : 0.00)$, $(S_{21} : \mathbf{0.50})$ | $(S_{22} : \mathbf{0.38})$ | $(S_{19} : 0.00)$ | $(S_{16} : \mathbf{0.25})$ | |

表 6 読書間隔とあらすじ文の位置の関係（4 回目読書前のあらすじ）

| 小説 ID | 1 日未満 | 1 日以上 3 日未満 | 3 日以上 7 日未満 | 7 日以上 14 日未満 | 14 日以上 30 日未満 | 30 日以上 |
|---|---|---|---|---|---|---|
| $N_1$ | $(S_5 : \mathbf{0.50})$, $(S_6 : 0.00)$ | $(S_4 : 0.00)$ | $(S_3 : 0.00)$ | | | |
| $N_2$ | $(S_{13} : \mathbf{0.50})$, $(S_{14} : 0.00)$ | | $(S_8 : \mathbf{0.13})$, $(S_{12} : 0.00)$ | | $(S_9 : 0.00)$ | |
| $N_3$ | | | | | | |

1 回目の読書範囲を指す．表 3 に，これら 3 種類の傾向と被験者ごとのあらすじ文選択傾向を示す．表中では被験者を $S_i$ と示しており，$S_1$ から $S_7$ が $N_1$，$S_8$ から $S_{15}$ が $N_2$，$S_{16}$ から $S_{22}$ が $N_3$ それぞれの小説を用いた実験の被験者である．

あらすじとして提示される情報として，すべての被験者がしおり直前の読書範囲の情報を必要としたことから，あらすじには読書中断直前の内容が求められると考えられる．一方で，しおり直前の読書範囲のみの情報で十分とする被験者が多いのに対し，小説冒頭や他の既読範囲から情報を必要とする被験者もおり，被験者に応じたあらすじ生成の必要性が示唆された．また，あらすじ文選択位置パターンには作品の影響は薄いと推察される．

しおり直前の 1 文は読書再開後の続きの内容に繋がる情報を含んでいる可能性がある．読書中断直前までの内容を想起するためにしおり直前の 1 文が重要とされると仮定し，被験者 $S_i$ のしおり直前文のあらすじへの採択確率 $p_i$ を調査した．確率 $p_i$ は下式 (1) により算出される．

$$p_i = \frac{\text{しおり直前文を含むあらすじの数}}{\text{あらすじ作成回数}} \tag{1}$$

表 4 に，$p_i$ 値の区間毎に被験者 ID と $p_i$ 値を $(S_i{:}\, p_i)$ の形式で示す．

被験者の過半数はしおり直前文をあらすじに選択しておらず，しおり直前文はあらすじとして重要であるとはいえない結果となった．しかし，しおり直前文を選択する被験者も一定数おり，毎回しおり直前文を選択している被験者も 2 名いた．あらすじ直前文が重要であるか否かは，被験者によって異なることが示唆された．

### 3.3 読書間隔とあらすじ

1 章で述べたように，読書間の時間間隔が長い場合には既読内容を忘れてしまう場合がある．これより，読書間隔が長いほど前提となる最初の部分からの情報が必要とされると仮定し，あらすじにおける初回読書範囲から選択されたあらすじ文の割合と読書間隔の関連を調査した．3 回目読書前，4 回目読書前に提示されるべきあらすじにおける結果を，それぞれ表 5，表 6 に示す．結果は，被験者 $S_i$ とあらすじにおける初回読書範囲から選択されたあらすじ文の割合 $f_i$ のセット $(S_i : f_i)$ により示す．$f_i$ は下式 (2) で算出される．

$$f_i = \frac{\text{初回読書範囲から選択された文数}}{\text{あらすじの総文数}} \tag{2}$$

ここで，2 章で述べたようにあらすじの総文数は常に 8 である．表 6 において，3 回の読書で実験を終了した被験者の結果は記載していない．

表 5， 6 に示すように，読書間隔が 1 日未満や 1 日以上 3 日未満の場合でも，初回読書範囲の文が選択された．また，3 回目読書時には読書間隔が 30 日以上の場合に初回読書範囲の文が選択されなかった．これらより，読書間隔はあらすじに影響せず，読書間隔が長いほど前提となる最初の部分からの情報が必要とされるという仮説は否定された．

同一の被験者が 3 回目読書時に作成したあらすじと 4 回目読書時に作成したあらすじにおける初回読書範囲の文に着目した．4 回目読書時に作成したあらすじにおいて，初回読書範囲に含まれる文を選択した被験者は，3 回目読書時に作成したあらすじにおいても初回読書範囲に含まれる文を選択していた．一方，4 回目読書時に初回読書範囲の文を選択しなかった被験者は，3 回目読書時にも初回読書範囲の文を選択していなかった．また，この傾向については，小説による差異も見られなかった．これらより，小説冒頭部分の情報を必要とするか否かは，各被験者固有の傾向であることが示された．

### 3.4 あらすじ中の意味情報

表 7 に，あらすじとして選択された文に含まれる意味情報について，キャラクター名，人称，場所，台詞の含まれる文の数を示す．同表中のパーセンテージは，それぞれ小説全文，1 名以上に選択されたあらすじ文，複数名に選択されたあらすじ文の総文数に占める割合である．

表 7　小説全文とあらすじ文中の各意味情報を含む文数．ただし，括弧内には，全文数に対して各文数が占める割合を示す．

| | | キャラクター名 | | 人称 | | 場所 | | 台詞 | | 総文数 |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_1$ | 小説全文 | 708 | (39.8%) | 756 | (42.5%) | 135 | (7.6%) | 715 | (40.2%) | 1179 |
| | 1 名以上に選択されたあらすじ文 | 61 | (64.9%) | 65 | (69.1%) | 23 | (24.5%) | 26 | (27.7%) | 94 |
| | 複数名に選択されたあらすじ文 | 13 | (65.0%) | 17 | (85.0%) | 6 | (30.0%) | 3 | (15.0%) | 20 |
| $N_2$ | 小説全文 | 393 | (35.7%) | 138 | (12.5%) | 77 | (7.0%) | 579 | (52.6%) | 1100 |
| | 1 名以上に選択されたあらすじ文 | 62 | (55.9%) | 22 | (19.8%) | 22 | (19.8%) | 37 | (33.3%) | 111 |
| | 複数名に選択されたあらすじ文 | 22 | (62.9%) | 7 | (20.0%) | 8 | (22.9%) | 11 | (31.4%) | 35 |
| $N_3$ | 小説全文 | 232 | (48.3%) | 49 | (10.2%) | 53 | (11.0%) | 116 | (24.2%) | 480 |
| | 1 名以上に選択されたあらすじ文 | 41 | (78.8%) | 1 | (1.9%) | 7 | (13.5%) | 2 | (3.8%) | 52 |
| | 複数名に選択されたあらすじ文 | 20 | (80.0%) | 1 | (4.0%) | 4 | (16.0%) | 0 | (0.0%) | 25 |

まず，キャラクター名について考察する．3 作品すべてにおいて，小説全文中での出現割合に対して，1 名以上および複数名に選択されたあらすじ文中での出現割合は，20%以上高い結果となった．あらすじ生成の指標として，キャラクター名の活用が多くの読者に必要とされるあらすじ文の抽出に貢献することが示唆された．場所情報についても，小説全文中での出現割合に対して，1 名以上および複数名に選択されたあらすじ文における出現割合が高い値を示したことから，場所情報もあらすじ生成の指標として有効活用できる可能性がある．ただし，$N_1$ と $N_2$ においては 10%以上の差があるのに対し $N_3$ においては 2.5%の差のみであり，小説の内容によっては有効性に差異が存在する可能性も示唆された．人称については，$N_1$ と $N_2$ では小説全文中よりもあらすじ文中の方が高い出現割合を示しているが，$N_3$ ではあらすじ文中よりも小説全文中でより高い出現割合を示した．小説全文中の人称を含む文の割合は，$N_1$ : 42.5%，$N_2$ : 12.5%，$N_3$ : 10.2% であった．$N_1$ があらすじ文中で高い出現割合を示したのは，語り手である「私」が頻繁に出現し，またその語り手が重要人物であるためであり,「私」がキャラクター名として捉えられていたためと考えられる．台詞については，3 作品すべてにおいて，小説全文中の出現割合よりもあらすじ文中の出現割合の方が 10%以上低い結果となった．この結果を利用し，台詞の含まれる文の重要度を減算することであらすじ生成に活用できる可能性がある．

## 4　おわりに

本稿では，読書進度に応じたあらすじの自動生成を目的とした調査研究として，人手によるあらすじ作成実験を行い，作成されたあらすじの特徴について考察を行った．考察の結果，以下の知見が得られた；

- しおりが挟まれる位置は，読者や作品によらず物語の区切りとなる箇所に偏る
- あらすじは,「しおり直前から抽出」「しおり直前と小説冒頭から抽出」「既読範囲全体から網羅的に抽出」の 3 パターンで作成される
- しおりの直前の文があらすじに含まれるか否かは被験者の特性による
- 読書間隔とあらすじ文の抽出源の間には関係は見られない
- キャラクター名と場所情報はあらすじ生成において有効に働く可能性がある

今後は，同一の読者が異なる小説に対してあらすじを作成した場合の傾向を考察する．また，本稿で明らかになったあらすじ生成に有効な知見を基に，読者の好みにあったあらすじの自動生成手法について検討していく．このとき，小説以外の情報（例えば，画像や地図情報）を用いた読書のエンターテインメント性の拡張についても議論していく．

## 謝辞

本研究は一部，科学研究費若手研究 B#16K21482 の助成のもと行われた．

### 参考文献

[1] H. Mori, R. Yamanishi, Y. Nishihara, and J. Fukumoto: The difference of word importance before and after bookmark for novel abstract in each reading progress, Proc. of the 21th Intl' Conf. on KES, pp.1246–1253, 2017.

[2] D. Wang, and T. Li: Document update summarization using incremental hierarchical clustering, Proc. of the 19th ACM Intl' Conf. on Information and Knowledge Management, pp.279–288, 2010.

[3] 今野勇気，荒木健治：遺伝的アルゴリズムを用いた文書間類似度による小説要約システムの性能評価，ARG 第 10 回 WI2 研究会予稿集，pp.19–20，2017.

[4] 野崎広志，中澤俊哉，重永実:物語理解におけるエピソード・ネットワークの構築,情報処理学会論文誌，vol.30，no.9，pp.1103–1110，1989.

[5] 白鳥裕士，中村聡史：スポーツジャンルに応ずるネタバレ特性分析と判定手法の提案, DEIM2016，2016.

[6] 田島一樹，中村聡史 : Twitter におけるアニメのネタバレツイート判定手法の提案, DEIM 2016，2016.

[7] 前田恭佑，土方嘉徳，中村聡史 : ストーリー文書を用いたレビュー文書でのネタバレ検出に関する一検討, 2016 年度 人工知能学会全国大会，2016.