ARG WI2 No.9, 2016
WI2-2016-25

しおりの前後での単語重要度の増加率を用いた小説の既読部分からのあらすじ生成

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

† 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　小説の読書再開時に，前回までの読書内容を忘れてしまい，読み返す場合がある．既読部分の読み返しにより内容を思い出し，スムーズに続きを読み始められる一方，本来の読書時間が削られてしまう．本稿では，読書中断地点を示すしおりの前後での単語の頻度変化に着目した既読部分からのあらすじ生成手法を提案する．具体的には，しおり以前から以後にかけての各単語の頻度の増加率を，既読部分のあらすじ生成における単語の重要度として扱う．この単語の重要度に基づいて，既読部分の各文の重要度を算出し，より重要度の高い文をあらすじとして抽出する．本稿では，あらすじ生成過程における，しおりの前後セクションの取り出し方法・基準や，あらすじ文の抽出対象範囲について，比較・検討した．また，人手でのあらすじ生成実験の結果と比較し，提案手法によって生成した既読部分のあらすじについて考察した．

キーワード　あらすじ生成，要約文抽出範囲，要約文抽出基準

1　はじめに

読書は，漫画や映画といった映像を伴うメディアに比較してより高い想像力を必要とする．漫画や映画では，ほぼ全ての場面に絵や映像，場合によっては音声が存在するため，想像を必要とせずに様子や情景を捉えることが容易である．一方で，読書は絵や映像が付与されていないことが多く，物語中の様子や情景は，読者の想像力によってのみ描かれる．情景を緻密に描くためには，物語の内容を充分に記憶し，把握する必要がある．しかし，長編小説の読書では「しおり」を挟み，読書を中断・再開することが多い．読書の再開時，物語の過去の出来事や内容を記憶していなければ，それまでの情景を描くことが困難であるのに加え，その後の情景を想像することも困難となっていく．

週刊連載等の漫画や小説では，閲読の中断によって遮断される記憶に対して，これまでの「あらすじ」によって補完する試みがなされている．前回の読書までの内容があらすじとして提供されることで，読者は読書中断前の重要箇所を確認可能となる．これらの週刊連載等のあらすじは，物語が進むに連れて変化していく．序盤では物語の設定や背景，人物紹介などが多く書かれる一方で，中盤以降では直前の出来事などが詳細に記述される．つまり，あらすじとして適当な文集合は読書の進度に応じて逐次的に変化していく．

本研究の最終目標は，ユーザの読書進度に応じて既読部分から自動的にあらすじを生成することにある．ここで，本稿におけるあらすじとは,「ユーザが続きを読む際に必要な情報」と定義する．未読部分の内容に直接関係のない既読内容はあらすじとしての価値が低い一方で，未読部分の内容に直接関連する箇所は読書再開後の情景の想像力を大きく左右する．本稿では，読書再開後に重要となる内容を重点的に評価し，既読部分からあらすじを自動生成する手法を提案する．

2　関連研究

あらすじ生成に関連した研究として，野崎らは，エピソード・ネットワークから物語のあらすじを作成する物語理解システムを構築している [1]．野崎らは物語全体の内容把握を目的としているが，本研究は既読部分の内容把握を目的としており，内容把握すべき範囲が異なる．特に，ユーザの読書進度に応じたダイナミックな変化を捉えることが本研究におけるキーポイントとなる．

既読内容の振り返りに関しては，岡田らの質問応答による読書支援の研究がある [2]．岡田らは，読書中に感じた疑問を質問応答によって解決することを試みている．既読部分の内容を忘れてしまった場合に読み返しの手間を省くという目的は本研究と共通しているが，既読部分の内容を想起するタイミングが異なる．岡田らの提案手法では，読書中に既読部分の内容を確認するため，その都度読書が中断されてしまう．一方，本研究で自動生成を目指す既読部分のあらすじは，読書再開前に提示可能であり，ユーザは事前に内容を想起できるため読書再開後は読書に集中することができる．

また，小説や映画，スポーツの勝敗にまつわるネタバレの防止に関連した研究も多数行われている．ネタバレとは，小説等のアイテムの楽しみを減らしうる記述のことを指す．小説においては，その多くが物語の結末や詳細なストーリー展開に関するものであるため，ネタバレ

情報は物語における重要情報であると言える．中村らの研究では，ネタバレ情報遮断のための情報曖昧化手法についての検討がなされている [3]．ここでは，ユーザが登録した関連キーワードによってネタバレ対象を登録し，正規表現によってネタバレ情報を検出している．中村らの研究における対象コンテンツはスポーツであるため，正規表現による勝敗や結果の判定が有効である．一方，本研究における対象コンテンツはジャンルを特定しない小説であり，小説に対して正規表現で重要情報を取得することは難しい．また，岩井らは，ネタバレ防止を目的としたあらすじ検出の研究を行っている [4]．あらすじか否かは，レビュー文書中の単語を素性とした機械学習に加え，レビュー文書の文書構造を用いることで判定している．レビュー文書中ではユーザの意見とあらすじが混在するため，それらの特徴の違いから機械学習によるあらすじ判定が有効と述べている．しかし，あらすじ生成においては小説中の全ての文があらすじとして採択される可能性をもち，その記述形式も共通の特徴をもつと考えられるため，記述の特徴の違いに着目しての機械学習は有効に働くとは考えがたい．

あらすじ生成において重要となる単語の重要度に関して，白鳥ら [5] や田島ら [6] の研究結果から，小説ジャンルごとのモデルを用いた機械学習による各ジャンルにおける重要単語判定の有効性が期待される．しかし，小説のジャンル数が多いことが課題となる．青空文庫 [7] では，国家別の分類や児童文学も含めると，約 60 種のジャンルが存在する．それらの中には属する小説が 1 作品のみのジャンルも存在し，学習データ数の問題により現状ではジャンルごとのモデル作成は現実的ではない．前田らは，小説テキストにおけるネタバレ単語の出現頻度・出現位置に着目している [8, 9]．具体的には，小説テキスト全体を 8 セクションに分割し，各セクションでのネタバレ単語の出現割合の変化を調査している．その結果，ネタバレ単語は物語後半に分布が偏るとされている．本稿では，前田らの調査結果に着想を得て，小説の既読部分からのあらすじ生成における単語重要度をデザインする．

3 提案手法

2 章で述べたように，ネタバレ情報ないしネタバレ単語は，物語における重要情報の部分集合であると言える．本研究では，まず，読書中断地点を示すしおりの前後それぞれのテキストからなる 2 セクションを作成する．ここで，前田らの調査結果では，物語中のネタバレ単語は全 8 セクションのうち後半 4 セクションに偏ることが確認された [8]．このことから，しおり以後の内容理解のための重要単語は，しおり以前・しおり以後の 2 セクションのうち，しおり以後に偏ると考えた．本研究では，この考え方をもとに，しおり以前から以後にかけての各単語の出現割合の増加率を，既読部分のあらすじ生成における単語の重要度として扱う．この単語の重要度に基づいて，既読部分の各文の重要度を算出し，より重要度の高い文をあらすじとして抽出する．このとき，単語重要度計算はネタバレ単語の大半を占める名詞・動詞・未知語に対してのみ行う．小説においては造語や人物名といった固有名詞が多く，それらは小説における重要単語であると考えられることから，未知語に対しても単語重要度を算出することとする．

また，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する．これは，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，あらすじが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．小説テキストには，青空文庫 [7] で公開されているテキストデータを用いる．このとき，あらかじめタイトル・著者名・ルビ・脚注等を削除したデータを用いる．

3.1 単語および文の重要度算出

まず，ユーザの読書記録から，ユーザの平均読書量を算出する．その後，しおり位置を基点として，しおり以前セクション（以下，セクション p）としおり以後セクション（以下，セクション f）を取り出す．各セクションの取り出し方法については，いくつかのバリエーションが考えられるため，後述の 3.2 節で詳細を説明する．

次に，各単語 x について，セクション p での単語出現頻度と，セクション f での tf 値をそれぞれ算出する．これらそれぞれを，各セクション内の単語種類数で除算することで正規化した値を $tf_p(x)$，$tf_f(x)$ とする．

得られた $tf_p(x)$，$tf_f(x)$ を用いて，既読部分のあらすじ生成のための単語重要度 $IW(x)$ を下式で算出する．

$$IW(x) = \log_2 \left(\frac{tf_f(x) + 1}{tf_p(x) + 1} \right). \tag{1}$$

文 i に含まれる単語集合を $\mathbf{S_i}$ としたとき，$\mathbf{S_i}$ に含まれる単語 x_j それぞれの重要度を合計したものを文 i の重要度 $IS(i)$ とする．$IS(i)$ は下式で求められる．

$$IS(i) = \sum_{j}^{N_i} WI(x_j). \tag{2}$$

ここで，x_j は $x_j \in \mathbf{S_i}$ を満たすものとし，N_i は文 i 中の重要度計算対象となる総単語数を示す．

3.2 セクション取り出し方法

しおりの前後セクションの取り出し方法・基準について比較する．これは，ユーザの読書進度を捉えるために

表 1 セクション取り出しパターン

パターン番号	セクション抽出基準	セクション p の評価指標
1	読書文数	実際の読書量
2		平均読書量
3	読書段落数	実際の読書量
4		平均読書量
5	読書単語数を満たす文	実際の読書量
6		平均読書量
7	読書単語数を満たす段落	実際の読書量
8		平均読書量

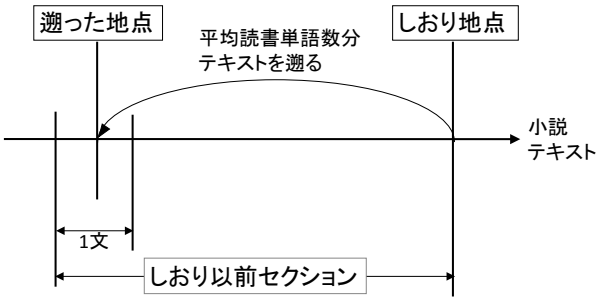

図 1 セクション抽出の概略図

適切なセクション範囲を検討するためである．表 1 にセクション抽出パターンを示す．抽出する際の評価指標，および抽出指標が異なる全部で 8 パターン用意した．

セクション p を抽出する際の指標として，表 1 中のパターン 1，3，5，7 では実際の前回読書量を用いる．パターン 2，4，6，8 については，平均読書量に基づいて抽出を行う．平均読書量は，1 回の読書で読む，文数，段落数，または単語数によって表現される．

次に抽出基準としては，文数，段落数，単語数を満たす文，単語数を満たす段落という 4 種類を用意した．パターン 1，2，3，4 では，しおり地点から平均読書文数，段落数をそれぞれ遡り，その地点からしおり地点までの範囲を，セクション p とする．パターン 5，6 では，図 1 に示す概略図のように，しおり地点から平均読書単語数を遡り，その地点を含む文，および，その地点からしおり地点までの文の集合をセクション p とする．パターン 7，8 では，パターン 5，6 と同様に，しおり地点から平均読書単語数を遡り，その地点を含む段落と，その地点からしおり地点までの段落の集合をセクション p とする．セクション f は，しおり地点からそれぞれ指定単語数・文数・段落数を進んだうえでセクション p 作成時と同様の方法で抽出する．

これら，4 種類の抽出基準と 2 種類の評価指標の組み合わせによる計 8 パターンのセクション抽出を行う．それぞれのパターンでの出力結果について比較・検討する．

3.3 抽出対象とするテキスト範囲

あらすじを出力するためのテキストの抽出対象範囲として，以下の 2 つが考えられる．

(1) セクション p から出力する場合

(2) 既読範囲すべてから出力する場合

(1) の場合，出力される内容は直前の内容であるため，続きの内容に直接関連する内容が提示される可能性が高く，読書支援として適しているといえる．しかし，直前の内容のみしか出力されないため，物語全体を通した話の流れが含まれない．また，冒頭部分に多く含まれる，登場人物に関する情報や物語の背景などの情報も含まれない．さらに，しおり位置が場面転換を挟む位置であった場合，続きの内容に直接関連しない内容のみがあらすじとして出力されてしまう．

一方，(2) の場合，すべての内容から抽出されるため，物語全体を通した話の流れや冒頭の情報が含まれる可能性がある．特に，場面転換時には直前の内容よりもより読書再開後の内容に関連する文が出力されることが期待される．しかし，多くの場合に直前の話の流れを含む小説の続きに直接関連した内容を含まないあらすじとなってしまう可能性がある．

あらすじを生成するための文の抽出源について，これら 2 種類それぞれを対象として，あらすじを生成実験を行う．それぞれの手法の利点・不利点について考察を行う．

4 あらすじ生成実験

あらすじ生成は，3.2 節で述べたセクション取り出しについての 8 パターン，3.3 節で述べたあらすじ文の抽出対象とするテキスト範囲 2 パターンの組み合わせにより，全部で 1 6 パターンの手法それぞれについて行った．各手法で生成されたあらすじについて，考察する．

4.1 実験条件

本実験では，青空文庫 [7] において公開中の小説「モルグ街の殺人事件」 [10] のテキストデータを用いる[1]．表 2 に，本小説の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・脚注を削除した後，本文に含まれる数である．本小説の本文は 33,233 字であり，3 章で述べた本研究で対象とする 20,000 字以上の小説という条件を満たしている．文体が現代的であるため内容理解が比較的容易であり，2016 年 9 月の青空文庫アクセスランキングにおいて青空文庫で公開中の 13,858 作品[2]中，テキスト版アクセスランキング 75 位と高順位にある作

[1] 本小説は，前田らの研究 [8] でも実験に用いられていた．
[2] 2016.10.20 現在

表 2 実験に使用する小説

著者	題名	文字数	単語数	文数	段落数
エドガー アラン・ポー	モルグ街の殺人事件	33,232	21,086	874	84

表 3 実験に使用するテキスト

	開始位置		終了位置		文字数	単語数	文数	段落数
	段落番号	冒頭 10 文字	段落番号	末尾 10 文字				
テキスト 1	1	サイレーンがどんな歌	48	要するにだ、僕はこの	10,299	10,299	423	48
テキスト 2	49	私はびっくりして黙っ	59	そして、そのとき窓が	3,898	3,898	156	11

品である．これらのことから，本小説は読者数が多く，あらすじ生成の意義がある対象と考えた．ただし，本実験では「モルグ街の殺人事件」から表 3 の条件で取り出したテキストを使用し，それぞれテキスト 1，テキスト 2 とする．このテキスト 1 とテキスト 2 の間にしおりが挟まれたものとして，既読部分からのあらすじ生成実験を行う．

なお，あらすじとしての出力文数は 8 文とし，$IS(i)$ の値が高い上位 8 文をあらすじとして選出する．この文数は，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．

4.2 あらすじ生成結果

セクション p からあらすじ抽出した場合の 8 パターンの出力結果それぞれにおいて，$IS(i)$ の値が高い上位 3 文を表 4 に示す．パターン 1 と 2，パターン 3 と 4，パターン 5 と 6，パターン 7 と 8 のそれぞれの組み合わせでは，重要度の上位 3 文は同一文であった．重要度が 4 番目以降の文について見ると，パターン 1 と 2 では上位 6 番目，パターン 3 と 4 では上位 5 番目までが同一文であり，それ以降の文についても，異なる文は 1 文のみであった．また，パターン 5 と 6，パターン 7 と 8 については，選択された文すべてが同じ文，かつ重要度順でも同一であった．この結果から，各文の $IS(i)$ の相関関係については，セクション p の評価指標の違いは影響せず，セクションの抽出基準のみが寄与することが確認された．

また，既読範囲すべてからあらすじ抽出した場合，8 パターンの出力結果すべてにおいて，同一の 8 文があらすじとして出力された．既読範囲に含まれる全ての文を評価した場合には，各文の相対的な重要度は表 1 のパターンの違いに左右されないことが確認された．

セクション p からあらすじ抽出した場合には，8 パターンすべてで出力に含まれた文が 4 文あった．このうち 3 文には「殺人事件」「ル・ボン（容疑者名）」「警視総監」といった単語がそれぞれ含まれており，本作品の本筋である殺人事件に関連する内容であると考えられる．他の 1 文には殺人事件に関連しうる単語は含まれておらず，本作品の本筋には直接関係しない内容であると判断可能であった．また，何れかのパターンによって抽出された文の中には，「証人」「調査」といった，本作品の本筋に関連する単語が含まれる文が存在した．一方，既読範囲すべてからあらすじを抽出した場合，殺人事件に関連しうる「証拠」という単語が含まれた文は 1 文のみであり，他の 7 文には本筋（殺人事件）に直接関係する単語は含まれていなかった．この結果から，あらすじの抽出対象とするテキスト範囲として，セクション p を用いた方が，物語の本筋を捉えたあらすじが生成されることが示唆された．ただし，これはしおりを挟む位置によって結果が変動する可能性もあるため，さらなる追加実験および考察が必要となる．

5 人手で作成したあらすじとの比較・考察

本研究は文書要約の研究の一種とも捉えられる．要約文書の評価方法としては，n-gram に基づく自動評価方法が知られている [11] が，本稿では提案手法のセクション取り出しパターンについての検討を行うため，人間が作成したあらすじと比較して考察する．

5.1 人手でのあらすじ生成実験

「モルグ街の殺人事件」の内容を知らない 10 代後半から 20 代の男女 5 名に「モルグ街の殺人事件」を閲読させ，あらすじを作成させた．まず，表 3 中のテキスト 1 を閲読させる．テキスト 1 を読み終わった後，10 分間の休憩を挟む．この休憩は，実環境における「読書の中断時間」を擬似的に再現するためのものである．休憩の間，被験者には「モルグ街の殺人事件」に関する情報（本文やあらすじなど）を得ないよう注意し，それ以外は自由に過ごしてもらう．次に，テキスト 2 を閲読させる．この後，小説の内容を把握しているかを確認するため，テキスト 1，テキスト 2 の内容確認のための質問に答えてもらう．最後に被験者にあらすじを作成させる．あらすじに採用する文数は，4 章で示した提案手法の抽出文数 8 文とする．

前提として，以下の 2 点を伝える．

表 4 システムによる各パターンにおける出力文の一部

パターン番号	あらすじ文
1, 2	我々がいまやっているような調査では、『どんなことが起ったか』ということよりも、『在ったことのなかで、いままでにまったく起ったことのないのはどんなことか』と尋ねなければならない。
	部屋がひどく乱雑になっていたこと、死体が頭を下にして煙突のなかに突き上げてあったこと、老夫人の体がむごたらしく切りさいなまれていたこと、などの事実や、さっき言ったこと、それから僕がわざわざ言うまでもない他の事実などは、警察ご自慢の明敏さを完全に参らせてしまって、力をすっかり麻痺させてしまったのだね。
	星をちらりと見ることが——網膜の外側を（そこは内側よりも弱い光線を感じやすいのだ）星の方へ向けて横目で見ることが、星をはっきり見ることになる、——星の輝きがいちばんよくわかるのだ。
3, 4	星をちらりと見ることが——網膜の外側を（そこは内側よりも弱い光線を感じやすいのだ）星の方へ向けて横目で見ることが、星をはっきり見ることになる、——星の輝きがいちばんよくわかるのだ。
	僕は警視総監のG——を知っているから、必要な許可をとるのは簡単だろう」
	〔楽しみというのはこんな場合に用いるには妙な言葉だと思ったが、私は何も言わなかった〕それにまた、ル・ボンには前に世話になったことがあって、僕はその恩を忘れてはいない。
5, 6	我々がいまやっているような調査では、『どんなことが起ったか』ということよりも、『在ったことのなかで、いままでにまったく起ったことのないのはどんなことか』と尋ねなければならない。
	部屋がひどく乱雑になっていたこと、死体が頭を下にして煙突のなかに突き上げてあったこと、老夫人の体がむごたらしく切りさいなまれていたこと、などの事実や、さっき言ったこと、それから僕がわざわざ言うまでもない他の事実などは、警察ご自慢の明敏さを完全に参らせてしまって、力をすっかり麻痺させてしまったのだね。
	星をちらりと見ることが——網膜の外側を（そこは内側よりも弱い光線を感じやすいのだ）星の方へ向けて横目で見ることが、星をはっきり見ることになる、——星の輝きがいちばんよくわかるのだ。
7, 8	鋭い声というのは、この証人はイタリア人の声だと思っている。
	星をちらりと見ることが——網膜の外側を（そこは内側よりも弱い光線を感じやすいのだ）星の方へ向けて横目で見ることが、星をはっきり見ることになる、——星の輝きがいちばんよくわかるのだ。
	それは誰か一人の（あるいは数人の）人のはげしい苦悶の叫び声らしく、——大声で長くて、短い早口ではなかった。

- 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，テキスト 2 の読書直前に読むもので，テキスト 1 から抜き出して作成する.

内容確認のための質問に対して，被験者 5 人すべてが 90%以上の正答率を示した．全ての被験者が小説の内容を理解したうえであらすじを作成したと考え，被験者 5 人の作成したあらすじ文すべてを，有効な参考情報として扱う.

被験者によって選択された文の種類数は，27 文あった．このうち，複数の被験者に選択された文は，7 文であった．表 5 に，被験者に選択されたあらすじ文のうち，4 節でのあらすじ生成実験においてセクション p をあらすじ抽出対象とした場合の提案手法の出力と合致した文を示す．また，表 6 に，提案手法の各パターンでの出力について被験者が選択した文と合致した文数を示す.

5.2 考察

表 6 から，段落をセクション抽出基準としたパターン 3，4，7，8 に対して，文を基準としたパターン 1，2，5，6 は被験者によって選択された文との重複文数が僅かに多い結果となった．文単位で読書セクションを抽出した場合，段落単位で抽出した場合よりも適切なあらすじを出力できる可能性が示唆される．また，今回の結果では，「平均読書文数」と「平均読書単語数をみたす文」の間，「平均読書段落数」と「平均読書単語数をみたす段落」の間の差異はほとんど見られず，基準が文であるか段落であるかが差異の大きな要因であったと考察される.

観点を変えて，どのような文が人手であらすじとして採択されたのかについて考察する．表 7 に，被験者によって選択された文のテキスト 1 中の出現位置を示す．表中では，テキスト 1 を 3 分割した場合のそれぞれの領域に含まれる採択された文数を示している．分割後の各領域には，それぞれセクション p とおおよそ同等量の文量が含まれる．2 名以上に採択された文は各領域にほぼ均等に位置している一方で，何れかの 1 名に選択された被選択文すべてについては冒頭よりもテキスト 1 終了直前の領域に偏って位置していることが見て取れる．また，あらすじに出力するテキストの対象範囲を既読範囲すべてのテキストとした場合，人手で作成されたあらすじと合致した文は存在しなかった．対して，出力するテキストの対象範囲をセクション p とした場合，被験者のうちいずれか 1 名が採択した文は，表 6 に示すように，

表 5 提案手法と人手でのあらすじ生成実験において共通して得られたあらすじ文

出力に文が含まれたパターン番号	あらすじ文
1, 2, 3, 4, 5, 6, 7, 8	この殺人事件について言えばだ、それについての僕たちの意見を立てる前に、僕たち自身で少し調べてみようじゃないか。
2	また、彼らは、あの争っているように聞えた声と、階上には殺されたレスパネエ嬢のほかに誰も見あたらず、また階段をのぼってゆく一行の者に気づかれないで逃げる手段がないという事実との、辻褄を合わせることができないことでも途方に暮れている。
1, 2, 5, 6	部屋がひどく乱雑になっていたこと、死体が頭を下にして煙突のなかに突き上げてあったこと、老夫人の体がむごたらしく切りさいなまれていたこと、などの事実や、さっき言ったこと、それから僕がわざわざ言うまでもない他の事実などは、警察ご自慢の明敏さを完全に参らせてしまって、力をすっかり麻痺させてしまったのだね。

表 6　提案手法と人手で作成したあらすじとで一致した文数

パターン番号	人手によるあらすじ文と一致した文数
2	3
1, 5, 6	2
3, 4, 7, 8	1

表 7　実験によって得られた文のテキスト 1 の各位置における文数

テキスト範囲	冒頭 1/3	中間 1/3	末尾 1/3
いずれか 1 名が採択	9	15	16
2 名以上が採択	2	3	2

8 パターンすべての場合で 1 〜 3 文存在した．以上から，既読範囲すべてを出力対象とするよりもセクション p を抽出対象とする方が，再現率を高くあらすじを生成可能であることが示唆された．

6　おわりに

本稿では，しおり以前から以後にかけての単語出現割合の増加率を単語重要度とした，既読部分のあらすじ生成手法を提案した．また，あらすじ生成実験や人手で作成したあらすじとの比較を通して，テキスト抽出基準やあらすじの抽出範囲について考察した．

今後は，被験者を増やした上でのあらすじの調査や，$tf-idf$，$OkapiBM25$ といった他の単語重要度を用いたあらすじ生成との有用性の比較を行う．このとき，n-gram に基づく自動評価方法 [11] による客観的な評価も行っていく．人手によって作成されたあらすじでは，被験者 5 名中 4 名が共通して「メインキャラクターに関する情報」を含む 2 文を選択していた．今後は，出現頻度のみではなく，登場キャラクターや場所などの意味的な解釈も取り入れた情報への重み付けを適用していく．

謝辞

本研究は，一部，科研費若手 (B)#16K21482 の助成のもと行われた．記して謝意を示す．

参考文献

[1] 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.
[2] 岡田悟，荒川達也:質問応答技術を用いた小説読書支援システムの提案,知能と情報，vol.27，no.2，pp.608–615，2015.
[3] 中村聡史，小松孝徳：スポーツの勝敗にまつわるネタバレ防止手法：情報曖昧化の可能性,情報処理学会論文誌，vol.54，no.4，pp.1402–1412，2013.
[4] 岩井秀成，土方嘉徳，西田正吾：レビューの文脈一貫性を用いたあらすじ文判定手法,情報処理学会論文誌データベース，vol.7，no.2，pp.11–23，2014.
[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,第 8 回 データ工学と情報マネジメントに関するフォーラム，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,第 8 回 データ工学と情報マネジメントに関するフォーラム，2016.
[7] 青空文庫, http://www.aozora.gr.jp.
[8] 前田恭佑，土方嘉徳，中村聡史：ストーリー文書内のネタバレの記述に関する調査とレビュー文書でのネタバレ検出の試み,第 8 回 Web とデータベースに関するフォーラム，pp.32–39，2015.
[9] 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.
[10] ポー エドガー・アラン：モルグ街の殺人事件,: http://www.aozora.gr.jp/cards/000094/card605.html.
[11] C.Y. Lin, and E.H. Hovy: Automatic evaluation of summaries using n-gram co-occurrence statistics, Proc. of HLT NAACL, pp.71-78, 2003.