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

前田 恭佑[1,a)]　土方 嘉徳[1,b)]　中村 聡史[2,3,c)]

概要：Amazon.com や楽天市場などのショッピングサイトでは，商品やコンテンツ（以降，アイテム）に対してレビューを簡単に作成・閲覧することができる．小説や映画などのストーリーを持ったアイテムに対するレビューには，レビュアの感想や意見が存在する一方で，そのアイテムのストーリーに関する記述が存在する．その記述の中には，実際にアイテムを見た時の楽しみや感動が減ってしまう記述（本稿では「ネタバレ」と呼ぶ）が含まれる場合があり，問題である．本研究では，ネタバレがストーリーの進行における位置づけと関係があるのではないかという仮定のもとでネタバレの検出を行う．しかし，記述内容がストーリーの進行においてどの位置に対応するのかはレビュー文書単体からでは把握できない．これに対処するために，本研究では，レビュー文書とは別にアイテムのストーリーを記録した文書（以降，ストーリー文書）も用いる．本研究では，まずネタバレとストーリーの進行における位置づけとの関係を知るために，ストーリー文書内のネタバレに関する記述について調査を行う．調査で得られた結果を基に，実際のレビュー文書からのネタバレ検出の可能性について考察する．

1. はじめに

近年，一般の消費者が商品やコンテンツ（以降，まとめてアイテム）に対して，自分の意見や感想を Web 上で（一般には，ショッピングサイトや口コミサイトで）投稿することが盛んになりつつある．一般に，レビューはユーザの実体験に基づいて書かれているため，まだそのアイテムを購入していないユーザにとっては有益な情報となりうる．しかし，コミックや小説，映画などのストーリーを持ったアイテムに対するレビューには，レビュアの感想や意見のほかに，そのアイテムのストーリーに関する記述も存在する．その記述の中には，そのアイテムの結末や詳細なストーリーの展開に関する記述がある．例えば，推理小説で犯人の名前を挙げたり，トリックの内容を明かしたりすることなどが挙げられる．人は小説や映画を鑑賞するときには，次に何が起こるかを想像することを一つの楽しみとしている [1], [2]．そのため，上記のような記述を目にしてしまうと，ユーザは実際にアイテムを体験した時の感動や楽しみを減らしてしまう可能性がある [3]．また，近年では SNS を対象に，アイテムの楽しみを減らしうる発言を防止する試みもある [4]．このように，アイテムの楽しみを減らしうる記述は問題視されている．本稿では，このような記述をネタバレと呼ぶ．また，レビューについて書かれた文書（ユーザの投稿単位となる文書）をレビュー文書と呼ぶ．

本研究の目的は，レビュー文書からネタバレとなる記述を検出することである．これまで，ストーリーに関する記述を含むレビュー文書を検出する研究 [5] や，レビュー文書中からストーリーに関する記述を含む文を検出する研究 [6], [7] が行われている．しかし，ストーリーに関する記述のすべてが，ユーザの楽しみを削いでしまうとは限らない．多くのアイテムの公式サイトやショッピングサイトにあるアイテム紹介ページには，ユーザの興味を引かせるためのストーリーの導入部分に関する記述がある．このような記述は閲覧するユーザにとって有益なものといえる．一方で，結末が描かれるストーリーの終盤部分は，ユーザの多くが楽しみにしていると考えられる．そのため，ストーリーの終盤に関しての記述はユーザの楽しみを大きく減らしてしまう可能性が高い．

このように，ストーリーに関する記述であっても，ユーザへの影響の大きさ（楽しみを減らしてしまう程度）は，実際のストーリーの進行における位置づけにより大きく異なると考えられる．ストーリーの進行とは作品全体におけるストーリーの進み具合を指し，その位置づけとは例えば序盤・中盤・終盤のどの部分に当てはまるかを示す．作品に

[1] 大阪大学大学院 基礎工学研究科
[2] 明治大学 総合数理学部
[3] JST CREST
[a)] k.maeda@nishilab.sys.es.osaka-u.ac.jp
[b)] hijikata@sys.es.osaka-u.ac.jp
[c)] satoshi@snakamura.org

よっては，ストーリーの内容や作中のシーンの転換などにより，それぞれの部分の長さは異なったり，明確な分割が困難な場合もあるが，作品における大まかな章（部分）構成に相当すると考えている．従来の研究はレビュー文書のみからストーリーに関する記述や，ネタバレとなる記述を検出しようとしてきた．しかし，レビュー文書にはストーリーの進行に対応する情報が含まれていない．そのため，各記述がストーリーの進行においてどの位置にあるのかを把握することができなかった．

我々はこの問題に対処するために，レビュー文書とは別にアイテムのストーリーを記録した文書（ストーリー文書）を用いることを提案する．例えば，アイテムが小説であれば，その小説の全文や一部始終の要約文 [*1] などがこれに相当する．我々は，ストーリーの進行における位置を，ストーリー文書におけるテキスト位置（テキスト先頭からの文字数）で代用できるのではないかと考えた．つまり，ストーリー文書を使い，アイテムの内容に関する 1 つの記述（レビュー文書内の記述）がストーリーの進行においてどの位置にあるのかを，概ね対応付けることができると考えた．

しかし，ネタバレとなる記述がストーリー文書内のどの位置に出現する傾向があるのかは分かっていない．我々は，ネタバレがストーリー文書の後半部分と関係があると仮定することにした．多くの小説や映画では，その作品のクライマックスや感動する場面は，作品の後半で現れると考えたからである．そこで，ネタバレに関する記述（実際にはキーワード）がストーリー文書中でどのような分布で出現するのかを調査することにした．なお，ある記述（ストーリーの内容に関する記述）をネタバレであると思うかどうかは，ユーザにより異なると思われる [8]．この調査では，多くのユーザが重要なネタバレ（問題のあるネタバレ）と判定したものにのみ焦点を当て，調査を行う．

我々は，この調査のために，新たにデータセットを作成することにした．このデータセットは，小説を対象ドメインとして，複数人の評価者にいくつかの小説を読んでもらい，各自がネタバレと思う内容を自由記述で書いてもらったものである．多くの評価者が重要であると判断したネタバレについて，そのネタバレの内容を表現するのに必要な単語群を別の複数の評価者に選出してもらい，多くの評価者が選出した単語をネタバレに関連する単語とした．また，ストーリー文書には，小説本文を利用した．

最後に，上記の調査結果を基に，実際のレビュー文書からネタバレの検出を試みた．検出法には，ルールベースによる方法や機械学習による方法など様々なものが考えられるが，本研究ではまずはストーリー文書中で特定のパターンで（後半に偏って）出現する語が，どれだけネタバレを抽出する能力があるのかを知るために，まずはその語が用いられている箇所を定性的に分析することにした．

本稿の構成は以下のとおりである．2 章で関連研究について述べる．3 章でストーリー文書内のネタバレの記述に関する調査の手法について述べる．4 章でネタバレに関連する単語のデータセットの作成方法について述べる．5 章で調査の結果と考察について述べる．6 章で実際のレビュー文書を例にネタバレ検出の可能性について考察する．最後に 7 章でまとめを述べる．

2. 関連研究

この章では，レビュー文書からストーリーの内容に関する記述（以降，あらすじ）やストーリーに関するネタバレを検出した研究と，コミュニティを対象にしたスポーツイベントなどのネタバレを検出した研究について述べる．

2.1 レビュー文書からのあらすじ・ネタバレ検出の研究

インターネット上のレビューに対する研究はテキストマイニングの分野で広く行われている [9]．その中で，我々の研究に最も関連する分野は，ストーリーを伴うアイテムに対するレビュー文書に注目した研究分野である．この研究分野では，あらすじの検出を目的とした研究や，ネタバレの検出を目的とした研究が盛んに行われている．Guo らは，レビュー文書の文構造に着目して Latent Dirichlet Allocation（LDA）を利用することで，あらすじを文単位で検出している [5]．また，岩井らは，レビュー文書中の各文に対して，種々の機械学習のアルゴリズムでその文があらすじか否かの判定をし [6], [7]，あらすじ部分を黒塗りにして表示するシステムを提案した [8]．上記で紹介した研究はネタバレの検出でなく，あらすじの検出を行っている．しかし，あらすじにはユーザにとって有益な情報とネタバレの両方が含まれている．我々は，あらすじ全体ではなく，あらすじの中でもユーザに不利益を与えるネタバレを検出対象としている．

Boyd-Graber らは，アイテムの内容に関する短文を TV Tropes という筋書き共有サイト [*2] から収集し，短文中にネタバレらしい単語があるか否かを機械学習により判定している [10]．彼らは，レビュー文書から得られる単語と文構造を利用している．それに対して我々は，ストーリー文書を用いることで，レビュー中の記述がストーリー文書中のどの位置に存在するかを知り，それをネタバレ検出に応用しようとしている．

2.2 コミュニティ内でのネタバレ検出の研究

レビューサイトではなく，SNS 全体またはその一部のコミュニティ内でネタバレを遮断しようとする研究も存在す

[*1] Web 上には要約文を集合知として収集するサイトが存在する．例：http://hm-hm.net

[*2] 投稿される短文には，投稿者自身がその短文の内容をネタバレと思うかどうかのラベルも付与される http://tvtropes.org/pmwiki/pmwiki.php/Main/HomePage

る．Klein らはアイテムをどこまで視聴・閲覧したか（進行度）をユーザごとに記録しておき，進行度の早い人の発言にネタバレがあるかもしれないと注意喚起しようとしている [4]．ストーリーを持ったアイテム以外にも，実世界のイベントを対象にしたネタバレ検出の研究も存在する．Golbeck は，Twitter のタイムラインを対象にスポーツの結果をネタバレの対象として検出している [11]．Nakamura らは，実世界でイベントが行われる時間帯とユーザの行動する時間帯を考慮して，スポーツの結果をネタバレとして検出している [12]．これらの研究はスポーツの結果を対象としているが，我々はストーリーを持ったアイテムごとに，そのストーリーの内容に関するネタバレを検出することを目的にしている．

3. ストーリー文書内のネタバレの記述に関する調査の方法

本研究では，ネタバレの記述がストーリー文書中でどのように出現するか（出現位置の分布）を調査する．この章では，まずその調査の方針について述べる．次に，対象とするアイテムの種類とストーリー文書として用いるデータについて述べる．次に，文章を単語単位に分割するのに使用した形態素解析の処理について述べる．最後に調査手法について詳しく述べる．

3.1 調査の方針

本調査のために，我々はネタバレの記述を収集し，ネタバレに関する正解のデータセットを作成することにした．ネタバレの記述の収集には実際のレビュー文書から抜粋する方法も考えられる．しかし，本調査ではより多くの記述を収集するために，複数の評価者に決められたアイテムを閲覧してもらい，それに関するネタバレを箇条書きで記述してもらうことにした．この記述がストーリー文書中のどの位置に出現するのかを調査するのである．

しかしこの調査では，記述してもらった内容がストーリー文書中のどこに書かれているのかを特定する必要がある．しかし，入力してもらったテキストは評価者自身の言葉で書かれているため，文単位でテキストの完全一致により場所を特定することは困難である．そこで我々は，評価者が記述した文からその内容を表すのに必要な単語を抽出し（これをネタバレに関する単語のデータセットとした），単語単位で位置を特定することにした．記述された文を構成する単語が，ストーリー文書中でどのように分布するかが分かれば，その分布からその文のネタバレの可能性を推定できるかもしれない．

なお，ネタバレに対して不快に思う程度には個人差があると考えられる [8]．本稿では，多くの人が重要なネタバレと思う記述について顕著な傾向が得られるかを確かめる．そのため，ネタバレの文に対して，どれだけ多くの人がネタバレと判定するかという一般性と，どれだけ多くの人が深刻であると判定するかという重要性の両方の観点から段階付けを行う．この段階付けによって調査の対象となるネタバレの文を選択し，上記データセットを構築する．ネタバレに関連する単語のデータセット（以降，ネタバレ単語データセット）の作成方法については 4 章で詳しく述べる．

表 1 使用する小説

著者名	タイトル	ラベル	テキスト量
ドイル アーサー・コナン	赤毛連盟	item1	48KB
大阪 圭吉	デパートの絞刑吏	item2	25KB
宮沢 賢治	銀河鉄道の夜	item3	84KB
夏目 漱石	こころ	item4	366KB
ポー エドガー・アラン	モルグ街の殺人	item5	76KB

3.2 使用するアイテムとストーリー文書

ストーリーを持つアイテムには映画，小説，コミックなどさまざまな種類が存在する．その中で，我々は青空文庫 [*3] に掲載される小説を対象とした．理由は，アイテムのストーリー文書として，アイテムの本文の全文が利用できる上に，それをオンラインで簡単に入手できるためである．本研究では，青空文庫分の小説・物語カテゴリに属するアイテムから 5 つを選んだ（表 1 参照）．表 1 の右端の列は，ダウンロード時のテキスト量（KB）である．

本研究では青空文庫からダウンロードしたテキストファイルをストーリー文書に利用する．このテキストファイルにはストーリーに関係のない記述もあるため，それを排除する前処理について述べる．青空文庫からアイテムをダウンロードしたそのままの状態では，《》で表記されたルビ表現がある．具体例を挙げると，「下《おろ》して」のように記述されているのであるが，このままでは「おろ」という単語がノイズになってしまう．そのため，前処理としてルビ表現（《》書き）の部分を除去する．ほかにも，外字表現（※［］書き），章分け表現（［＃］書き），本文前後にあるタイトルや注釈などを除去する．これらは正規表現を利用することで完全に除去した．

3.3 形態素解析時の処理

形態素解析の処理は，ネタバレ単語データセットを作成する（4 章で述べる）際とストーリー文書内の単語の出現分布を分析する（次節で述べる）際に利用する．形態素解析には，MeCab[*4] を利用した．形態素解析によって得られる結果には，品詞と活用形，その原形も付与されている．しかし，得られる結果は参照する辞書に依存し，ストーリー文書に出現する人物名や独自の言葉（特有語）の多くは登録されていない．そのため，それらの単語の多くが期待通

[*3] 日本国内で，主に著作権の消滅した文学作品のテキストを公開している http://www.aozora.gr.jp/
[*4] オープンソースの形態素解析エンジン http://taku910.github.io/mecab/

りに抽出されない．そこで事前に形態素解析器の辞書への単語の追加登録を行った．辞書に登録をした単語はストーリー文書内に出現する名詞（人物名・特有語）と動詞で，まだ辞書に登録されていないものである．動詞は活用形を含めて登録した．また，活用形が使われた単語は原形の形に直して出力した．

なお，人物名は場面ごとに呼び方が異なる場合がある．例えば，「シャーロック・ホームズ」を「ホームズ」と呼んだり，フルネームで呼んだりする．今回は異なる呼び方でも同一の単語とみなすようにした（統一した呼称を原形に，それ以外の呼び名を活用形の一つとして辞書に登録した）．この辞書データはストーリー文書とネタバレの文に対して同じものを使用する．また，頻出する単語を除くために除去単語（ストップワード）を定義した．ストップワードのリストは SlothLib プロジェクト [*5] からダウンロードした．このリスト以外にも，「する」，「れる」，「られる」の動詞は多くの文で頻出するためストップワードとした．また，ひらがな一文字やカタカナ一文字も意味判別が困難な単語であるためストップワードとした．これらのストップワードは形態素解析器で結果を出力する際に除去され，本研究では使用されない．

3.4 出現分布の調査手法

ストーリー文書内で単語がどのような分布で出現するか（出現パターン）を分析する手法について述べる．我々は，ストーリー文書を文字数を基に均等に分割し（分割されたそれぞれの塊をパートと呼ぶ），各パートにおける出現割合（各パートにおける単語の出現回数 / 全パートにおける単語の出現回数）を単語ごとに求める．次に，出現割合を前半部から後半部へ順に足し合わせたもの（累積出現割合と呼ぶ）を単語ごとに求める．このパートごとに推移する累積出現割合をその単語の出現パターンとみなす．具体的に，今回はすべてのアイテムに対して 8 分割で分割をしている．8 分割の理由は，前半・後半と明確に分けられる 2 分割の累乗数の中で，分析しやすい分割数であったためである．我々は，それぞれのパートが作品における大まかな章に相当すると考えているため，大きな分割数は適当ではないと考えた．単語単位での調査を行うため，ストーリー文書を形態素解析にかけ，全単語の出現パターンを求める．

我々は，後半部分に注目した分析を行うため，後半 4 パートにおける累積出現割合の変化をみる．我々は 3 つの出現パターンを定義した．その概念図を図 1 に示す．1 つ目は前半に比べて後半での出現割合が大きいパターン（パターン 1）である．2 つ目はパターン 1 の中でも最後の 8 パート目で出現割合がちょうど 1 になるパターン（パターン 2）である．3 つ目は，ストーリーの最初から最後まで均等に

[*5] http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt

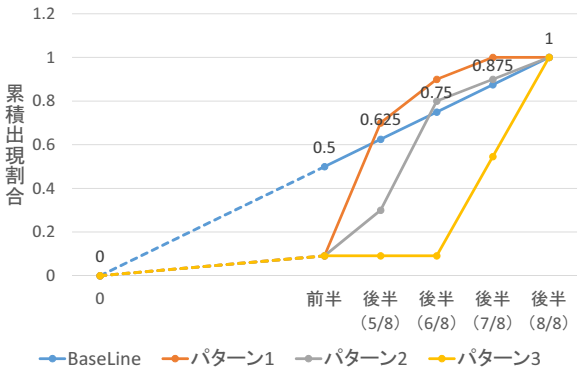

図 1 単語の出現パターン

出現する出現パターンを BaseLine と考え，累積出現割合が常に BaseLine を下回るパターン（パターン 3）である．これらのパターンは包含関係にある（パターン 1 ⊇ パターン 2 ⊇ パターン 3）．パターン 1 は，ネタバレの内容がストーリーの後半に偏っているという仮説をそのままパターンとして定義したものである．パターン 3 は作品の最終場面（クライマックス）において急増する単語を調べるために設定した．パターン 2 は，出現の仕方が上記 2 つの中間にあたるもので，後半に偏っており，なおかつ最終場面まで出現し続ける単語を調べるために設定した．

4. ネタバレ単語データセット

この章では，ネタバレ単語データセットの作成手順とその特徴について述べる．はじめに，作成手順の概要を述べ，次にネタバレの文とそれを構成するのに必要となる単語を得るために評価者に取り組んでもらうタスクについて述べる．最後に，データセットの特徴について述べる．

4.1 データセット作成手順の概要

ネタバレ単語データセットを作成するための評価者へのタスクについて説明する．ネタバレの文を記述するタスクを行う評価者は 6 人（男性 3 人，女性 3 人）で，平均年齢は 19.5 歳で，全員日本人の大学生である．タスクは全部で 3 つ（タスク 1 – 3）あり，全評価者がすべてのタスクを実行する．タスク 1 は 2014 年 12 月から 2015 年の 1 月にかけて行われた．タスク 2，タスク 3 は 2015 年 6 月から 7 月の間で行われた．ネタバレの文を構成する単語を抽出するタスクを行う評価者は男性 5 名で，平均年齢が 22.6 歳で，全員日本人の大学院生である．このグループのタスクは 1 つ（タスク 4）である．タスク 4 は 2015 年 8 月に行われた．以下の項で，各タスクの目的とその詳細を述べる．

4.1.1 タスク 1：小説の読書とネタバレ

タスク 1 は，ネタバレの記述を集めることと，それらの文の重要度を決めることを目的としている．評価者に表 1 で示した 5 つの小説を読んでもらい，そのアイテムのネタバレを記述してもらう．評価者にはネタバレを “これから

作品を読む人が聞いたら楽しみが減ってしまう内容” と説明した．ネタバレは箇条書きの短文で，思いつく限り書いてもらう．また，すべてのネタバレを記述した後に，それぞれのネタバレの文について 1 から 5 でネタバレ度合いをつけてもらう（1 – 少々のネタバレ，5 – 重要なネタバレ）．

ここで，ネタバレの記述を集めた際に，各文に対して以下の処理を行う．

1 誤字，脱字の修正

2 文頭の接続語，指示語の削除

1 つ目は筆者が判断をして修正をした．2 つ目は文頭にある「しかし」,「それから」といった接続語,「その」,「あの」といった指示語を除去する．評価者は，前の箇条書きの内容を受けて記述しているケースがあった．今後のタスクでは，文単位で（前後のつながりを無視して）評価してもらうため，これらの語は除去する．

4.1.2 タスク 2 : 自分の書いた文へのネタバレ度合いづけ

タスク 2 は，時間経過によるネタバレ度合いの変化があるかどうかをみることを目的としている．評価者に，再度（約半年の期間を空けて），自分の書いたネタバレの文に 1 から 5 でネタバレ度合いをつけてもらう（1 – 少々のネタバレ，5 – 重要なネタバレ）．タスク 1 のネタバレ度合いと比較して，評価者の基準が一定であったのかを判断する．

4.1.3 タスク 3 : 他人の書いた文へのネタバレ度合いづけ

タスク 3 は，他者からの評価も含めた信頼性の高いネタバレの記述を得ることを目的としている．評価者に，自分以外の 5 人が書いたネタバレの文に 0 から 5 でネタバレ度合いをつけてもらう（0 – ネタバレと思わない，1 – 少々のネタバレ，5 – 重要なネタバレ）．複数人によるネタバレ度合いを得ることで，各文のネタバレ度合いの一般化が可能となる．例えば，過半数の評価者が高いネタバレ度合いをつけている文を，大多数が重要と考えるネタバレとすることができる．今回の調査では，評価者の過半数（4 人以上）が 4 以上のネタバレ度合いをつけた文のみを使用する．

4.1.4 タスク 4 : ネタバレを構成するのに必要な文節の選択

タスク 4 は，ネタバレに関連する単語を各文から抽出することを目的としている．ネタバレを記述した評価者とは別の 5 人の評価者にタスクを行ってもらう．タスクの内容は，タスク 3 で選択された文の内容を表すのに必要な，最低限の数の文節を選ぶことである．このタスクは，文そのものが持つ意味についてのみ注目すれば良いので，アイテムの内容を知らなくても行えると判断した．文の文節分けには，日本語係り受け解析器の CaboCha[*6] を利用する．文節は “/” で区切る．元の文と，文節分けした文を提示し，文節を丸で囲むようにして選択させる．評価者の過半数（3 人以上）に選ばれた文節を収集する．収集した文節を形態素解析して，意味のある単語を抽出する．具体的には，名詞・動詞・形容詞・副詞を抽出した．これをネタバレ単語データセットとする．3 章で述べたとおり，人物名は統一した呼称に，その他の品詞も原形に直している．

[*6] http://taku910.github.io/cabocha/

4.2 ネタバレ単語データセットの特徴

各タスクの結果と得られたデータを示し，その特徴について説明する．

4.2.1 タスク 1

タスク 1 の結果を表 2 に示す．アイテムにおいて，記述された文の数の平均は 96.6 であった．どの評価者においても，小説の文量（表 1 参照）が大きいほど記述量も増す傾向があった．また，平均よりも多くの文を書いている評価者の文の内容はストーリー全体を網羅的に書いており，一文ごとの長さも長くなる傾向があった．

表 2 記述されたネタバレの文の数

	user1	user2	user3	user4	user5	user6	ALL
item1	15	12	17	6	11	11	72
item2	18	7	18	12	9	17	81
item3	34	19	31	17	7	14	122
item4	44	19	91	14	17	22	207
item5	15	17	29	11	13	13	98
ALL	126	74	186	60	57	77	580
1 文ごとの平均文字数	55.3	25.1	46.5	23.2	27.2	28.2	

4.2.2 タスク 2

評価者がタスク 1 の時につけたネタバレ度合いと，タスク 2 の時につけたネタバレ度合いを表 3 に示す．タスク 1 とタスク 2 で，ネタバレ度合いをつける評価基準に変化がなかったかを調べる．この調査にはエーベルの級内相関係数（ICC）[13] を用いた．ICC には，1 人の評価者が複数回評価した時の評価者内信頼性（ICC(1, 1)）と，複数の評価者が 1 回評価した時の評価者間信頼性（ICC(2, 1)）があり，それぞれの値が信頼性の指標とされる．今回の調査には ICC(1, 1) を用いる．ICC(1, 1) の値は表 3 に示している．Landis らの ICC の値の解釈 [14] をもとにすると，0.61 から 0.80 で概ね一致していると言える．

またネタバレ度合いの差（タスク 2 のネタバレ度合い – タスク 1 のネタバレ度合い）を算出した．この結果を図 2 に示す．差が –1, 0, 1 であった文の数が全体で約 85%を占めている．このことからネタバレ度合いの大幅な変更は少なかったと言える．以上のことから，ネタバレ度合いの基準は時間が経過してもほぼ一定であり，この数値が信頼できるものといえる．

4.2.3 タスク 3

タスク 3 の結果を基に，評価者の過半数（4 人以上）が 4 以上のネタバレ度合いをつけた文を特定する．その文の数は表 4 に示している（表 4 中の “対象となる文数”）．また，

表 3　タスク毎のネタバレ度合いと ICC

	タスク 1		タスク 2		
	平均値	分散	平均値	分散	ICC(1, 1)
user1	2.65	1.9	2.97	1.83	.601
user2	3.02	1.64	2.86	1.7	.710
user3	3.58	1.72	3.33	1.49	.713
user4	2.83	2.27	3.01	1.88	.760
user5	3	2.14	3.21	2.45	.795
user6	3.14	1.99	2.79	2.21	.692
ALL	3.11	1.99	3.08	1.85	.708

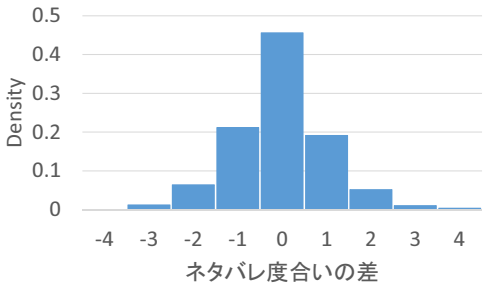

図 2　ネタバレ度合いの差のヒストグラム

同一の文に対する評価者全員のネタバレ度合いの一致度合いを評価者間信頼性を示す ICC(2,1) によって調べる．文を書いた本人の評価値はタスク 2 の時のものを利用した．これは他の評価者が評価した時期と同時期のものを用いるためである．ICC(2, 1) は.591 となった．[14] の解釈からは，これは中程度の一致といえる．このことから，文の内容がネタバレとして深刻であるかどうかの判断は，多少のズレはあるが評価者間で一致する傾向にあるといえる．

4.2.4　タスク 4

5 人の評価者による文節選択の結果について述べる．このタスクの結果，表 4 に示す数の単語が得られた（表 4 中の “抽出単語数”）．() の中の値は名詞と動詞の数である．重複して出現した単語は削除している．ここで得られた単語について定性的な結果を述べる．抽出した単語で名詞・動詞以外の単語はわずかに 4 つであった（「荒々しい」,「鋭い」,「ひとりでに」,「早い」）．このことから，名詞と動詞がネタバレに関連しやすい品詞であるといえる．これからのストーリー文書での分布の分析では，名詞と動詞に限定する．また，従来の研究 [5], [6], [7], [10] にも示されていたように，登場人物名やアイテムに特有な語がネタバレ単語データセットに幾つか含まれていた．

5.　ストーリー文書内のネタバレの記述に関する調査の結果

この章ではまず，ネタバレ単語データセットの単語でストーリー文書内に出現しない単語に対する前処理について述べる．次にネタバレ単語データセットがストーリー文書中でどのような分布で出現したかについての結果を示す．

表 4　過半数（4 人以上）が 4 以上のネタバレ度合いをつけた文の数と，その中の抽出された単語の数（抽出単語数）
抽出単語数の括弧内の数字は，抽出単語のうち名詞と動詞の数

	全文数	対象となる文数	抽出単語数
item1	73	25	24 (24)
item2	81	24	33 (33)
item3	122	25	35 (35)
item4	207	43	64 (63)
item5	98	24	69 (66)
ALL	581	141	225 (221)

5.1　ネタバレ単語データセットに対する前処理

ネタバレ単語データセットにはストーリー文書内には存在しない単語が幾つかみられた．これらは以下のように分類できる．

- 別の単語への言い換え
  例：強盗→盗む
- 漢字・送り仮名の違い
  例：睡る→眠る，諦らめる→諦める
- 動詞と名詞の違い
  例：死ぬ→死
- 評価者独自の表現（類似単語も出てこない）

このうち，「評価者独自の表現」とは，評価者がアイテム中に明確に記載されていないストーリーの展開を予想して書いている言葉である．例として，銀河鉄道の夜（item3）で書かれた次の文を挙げる．

カムパネルラは川でザネリを助けて溺れ死んでしまう

本文中では，溺れることまでは記述されているが，直接死んだことについては書かれていない．このように評価者がアイテムの内容を解釈してネタバレを書くこともある．今回，評価者独自の表現以外であれば，明確にわかる範囲内で筆者が本文中の単語に置き換えた．置き換えた単語数は全体で 20 個であった．

5.2　ネタバレ単語データセット中の単語の出現分布の分析

ネタバレ単語データセットとストーリー文書内の全単語に対して，パターン 1 から 3 に該当する単語の割合を比較することで，ネタバレ単語の分布の傾向を知る．ストーリー文書から抽出した単語数と，それぞれの出現パターンに該当する単語の割合を表 5 に示す．ネタバレ単語データセットについて，ストーリー文書内に存在する単語，それぞれのパターンに該当する単語の割合を表 6 に示す．また，例として「赤毛連盟（item1）」のネタバレ単語データセット中の単語の分布を図 3 に示し，定性的な分析を行う．図 3 は，ネタバレ単語データセット中の単語 24 個を 4 つのグラフに分けて示している（各グラフに BaseLine も記載）．

まず，表 5，表 6 をもとに定量的な分析を行う．多くのアイテムについて，ストーリー文書内の単語でパターン 1 に該当する単語の割合は半分以下である．一方，ネタバレ

表 5　ストーリー文書内の全単語数と各パターンの割合

	全単語数	パターン 1	パターン 2	パターン 3
item1	1702	0.514	0.162	0.142
item2	1084	0.408	0.145	0.131
item3	1637	0.415	0.138	0.113
item4	4629	0.433	0.175	0.140
item5	1884	0.388	0.171	0.152

表 6　ネタバレ単語データセットのうちストーリー文書内に存在する単語数と各パターンの割合

	単語数*	パターン 1	パターン 2	パターン 3
item1	20	0.8	0.65	0.6
item2	26	0.576	0.461	0.461
item3	25	0.68	0.44	0.24
item4	58	0.586	0.534	0.396
item5	51	0.647	0.411	0.294

* ネタバレ単語データセットのうちストーリー文書内に存在する単語数

単語データセットでパターン 1 に該当する単語の割合はネタバレ単語データセット（かつストーリー文書にも存在するもの）の半分以上であった．また，すべてのアイテムにおいて，ストーリー文書内の単語でパターン 2，パターン 3 に該当する単語の割合は 0.2 以下である．一方，ネタバレ単語データセットでパターン 2，パターン 3 に該当する単語の割合は，すべてのアイテムおいて 0.2 以上であり，そのうち多くは 0.4〜0.7 である．以上のことから，ネタバレの記述においては前半より後半に偏った単語が使用される傾向があるといえる．

次に定性的な分析を行う．ネタバレ単語データセット中に登場人物名や特有語はいくつかあるが，それらの分布は必ずしも後半に偏ったものではなかった．例えば，赤毛連盟では，「赤毛連盟（図 3 左下）」という単語は前半からほぼ均等に出現している．「ダンカン・ロス（図 3 左下）」，「スポールディング（図 3 右下）」といった登場人物名も，ネタバレ単語データセットに入っているが，ストーリー文書の後半にはほとんど出現していない． 作品の舞台や登場する団体名など作品を通して出てくる語は想定した分布とはならなかった．また，作品の前半で謎として与えられるような語も，想定した分布とはならなかった．

6. レビュー文書からのネタバレ検出

ストーリー文書中のパターン 2，パターン 3 の単語が含まれるレビューを探し，その内容について考察する．パターン 2 とパターン 3 に注目して探した理由は，パターン 1 は該当する単語が多いためである（表 5 参照）．実際のレビュー文書は Booklog[*7] から収集した．Booklog のレビューには，その内容がネタバレか否かを投稿者が判断して，そのラベルをつけることができる．ここでは例として，「モルグ街の殺人（item5）」の結果（下記「レビュー文書 1」と「レビュー文書 2」参照）ついて述べる．下線はパターン 2 に該当する単語，2 重下線はパターン 3 に該当する単語である．レビュー文書 1 にはネタバレのラベルが付与されていた．「動物」，「窓」といった作中のキーワードが存在する一方で，「思う」，「想像」といった実際にはあまり重要でない単語も存在した．レビュー文書 2 にはネタバレのラベルがつけられてはいなかったが，パターンに該当する単語が多くあった．この中で，「手」，「上」といった単語は，本文中とは異なる用途で使用されている．

実際のレビュー文書からパターンに当てはまる単語を探した結果，今回提案した手法は，作中のキーワードを直接使っているレビュー文書に対しては，ネタバレを検出できる可能性があることが分かった．また，パターンに当てはまる単語であっても，本文中と異なる使われ方（例えば，「禁じ手」，「斜め上」のように慣用句の一部になっている）がされる場合も見つかり，検出の正確性については課題が残ることも分かった．

レビュー文書 1（「モルグ街の殺人」）〈 ネタバレ 〉

やっと読めた！！！いつか読みたい，と 思い 本棚に登録して 1 年 8 か月が経ってたことに驚いた（笑）．以下ネタバレ（と自己満の感想）あります．モルグ街の 殺人…犯人がまさかの 動物 ！だから言語が一致しないワケだ．殺害現場が結構細かく描写されていて，想像 すると気持ち悪くなる．窓 の仕掛けはイマイチ分からず..

レビュー文書 2（「モルグ街の殺人」）

近代推理小説の始まりにして，かなりの完成度．往々にして，新しいジャンルの開拓者は，開拓時点でかなりの傑作を残すものですが，今作はその通り．
その記念碑的な，モルグ街の 殺人 事件．推理小説の始まりがまさかのオチで驚く．いきなり禁じ 手 に近いようなところ．そう考えると，推理小説とは，いかに読者を騙すか，というよりもいかに読者の想像の斜め 上 をいくか，というエンターテイメント性に本質があるのかもしれない．
推理を行う デュパン さんの 言う ことは難解．数理的な思考を超絶的な語彙力で，説明が説明になっていない．本質から入り，細部に 入りこみ，説教に移り，やっと謎解きに進むあたり，文学寄りなんでしょうね．

7. おわりに

本研究では，ストーリーをもつアイテムについて書かれたレビュー文書を対象に，ストーリーの進行における位置と対応付けてネタバレを検出することを提案した．ストーリーの進行の情報を把握するために，アイテムごとにストーリー文書を利用した．

ストーリー文書においてネタバレに関する記述を調査した結果，ネタバレに関連する単語は後半に偏って出現する

*7 http://booklog.jp/

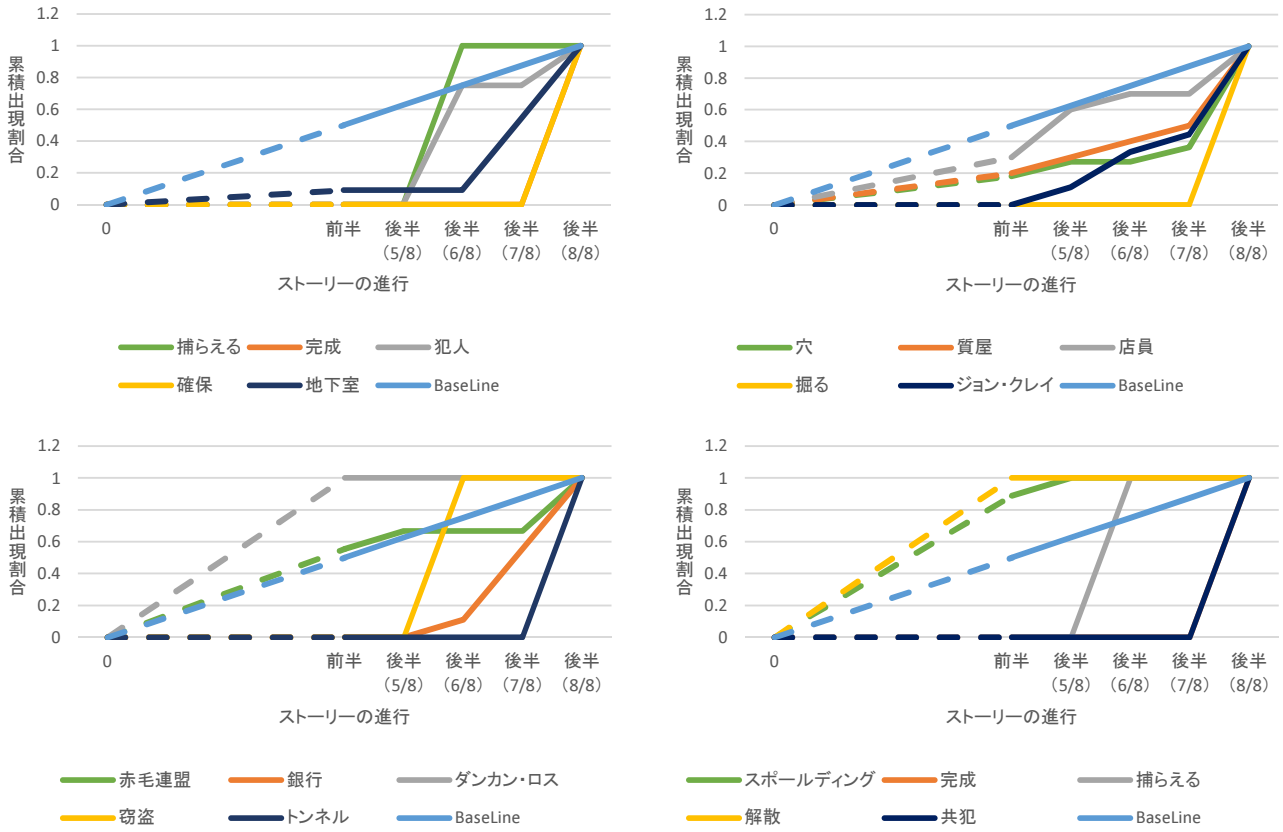

図 3 ネタバレ単語データセットの単語の分布（赤毛連盟）

傾向が見られた．また，実際のレビュー文書において，後半に偏って出現する単語の使われ方を定性的に分析したところ，提案した手法はレビュアの言い換えには対応出来ていないが，レビュアがストーリー文書中に出現する単語を用いてネタバレを記述していれば，それを検出できる可能性があることが分かった．

今後はストーリー文書から該当するパターンの単語を抽出したときに，別のアイテムのストーリー文書における分布と比較することで，よりネタバレに関連する単語のみに限定することや，レビュアの言い換えに対応することに取り組む予定である．

謝辞

本研究は日本学術振興会科学研究費補助金（課題番号：25540080）の助成を受けたものである．

参考文献

[1] Loewenstein, G.: The psychology of curiosity: A review and reinterpretation. Psychological bulletin, Vol.116, No.1, pp.75–98 (1994).

[2] Wilson, T., Centerbar, D., Kermer, D. and Gilbert, D.: The pleasures of uncertainty: prolonging positive moods in ways people do not anticipate, Journal of personality and social psychology, Vol.88, No.1, pp.5–21 (2005).

[3] Tsang A.S. and Yan, D.: Reducing the spoiler effect in experiential consumption, Advances in consumer research, Vol.36, pp.708–709 (2009).

[4] Klein, D. and Jackson D.: Processing content spoilers, U.S. Patent Application 20140101244 (2014–4–10).

[5] Guo, S. and Ramakrishnan, N.: Finding the storyteller: automatic spoiler tagging using linguistic cues. Proc. of COLING '10, pp.412–420 (2010).

[6] 岩井秀成，池田郁，土方嘉徳，西田正吾：レビュー文を対象としたあらすじ分類手法の提案，電子情報通信学会論文誌，Vol.J96-D, No.5, pp.1222–1234 (2013).

[7] 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 (TOD)，Vol.7, No.2, pp.11–23 (2014).

[8] 岩井秀成，池田郁，土方嘉徳，西田正吾：レビュー文を対象としたあらすじ分類手法の提案とあらすじ非表示システムの開発，インタラクション 2013 論文集，pp.1–8 (2013).

[9] Pang, B. and Lillian L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2.1–2: 1–135 (2008).

[10] Boyd-Graber, J., Glasgow, K. and Zajac, J.: Spoiler alert: machine learning approaches to detect social media posts with revelatory information. ASIST 2013, Vol.50, No.1, pp.1–9 (2013).

[11] Golbeck, J.: The twitter mute button: a web filtering challenge. Proc. of CHI '12, pp.2755–2758 (2012).

[12] Nakamura, S. and Tanaka, K.: Temporal filtering system to reduce the risk of spoiling a user's enjoyment. Proc. of IUI '07, pp.345–348 (2007).

[13] 対馬栄輝：信頼性指標としての級内相関係数，http://www.hs.hirosaki-u.ac.jp/ pteiki/research/stat/icc.pdf

[14] Landis, J.R. and Koch G.G.: The measurement of observer agreement for categorical data, Biometrics, pp.159–174 (1977).