3O4-OS-04b-2in1

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

A Basic Study on Spoiler Detection from Review Comments Using Story Documents

前田恭佑 *1　土方嘉徳 *1　中村聡史 *2
Kyosuke Maeda　Yoshinori Hijikata　Satoshi Nakamura

*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
Osaka University　Meiji University

In many shopping sites like Amazon.com, it is possible to view and create a review about the product (item). The review of some items including story, such as novels, comics and movies, contain a description of the contents of that story. In some cases, that description spoil the ending of a story. Blocking the spoilers, it may be possible to improve user satisfaction. In this study, we propose a detection method of spoilers using the story documents that record the whole story of each item.

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 上で投稿することが盛んになりつつある．しかし，小説や映画などのストーリーを持つアイテムに対するレビューには，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．このような記述は，ユーザが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 [1, 2, 3]．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 [4] 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 [5, 6] 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．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（実際には単語）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．

我々は，この調査のために，ネタバレに関連する単語のデータセットを作成することにした．まず，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う．次に，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う．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．

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．本研究で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，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．3 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．4 章で調査の結果と考察について述べる．5 章で調査の結果を基にしたネタバレ検出の手法について考察する．最後に 6 章でまとめを述べる．

2. 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

連絡先: 前田恭佑，大阪大学大学院 基礎工学研究科，
06-6850-6383，k.maeda@hlab.sys.es.osaka-u.ac.jp

*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

表 1: 使用する小説

著者名	タイトル	テキスト量
item1	赤毛連盟	48KB
item2	デパートの絞刑吏	25KB
item3	銀河鉄道の夜	84KB
item4	こころ	366KB
item5	モルグ街の殺人	76KB

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．最後に調査手法について詳しく述べる．

2.1 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，ネタバレのデータセットを作成することにした．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．

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可能性がある．

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．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 3 章で詳しく述べる．

2.2 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や小説，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 *2 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．

2.3 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），各パートにおける単語の出現割合（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．

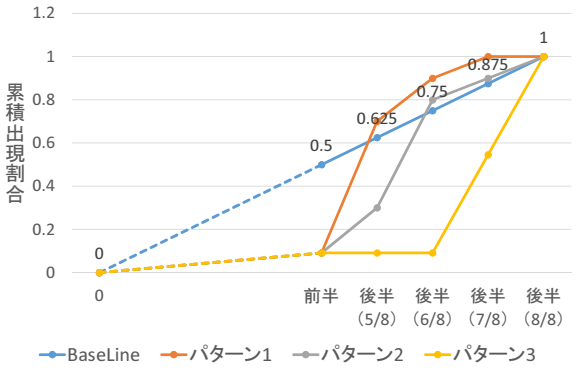

図 1: 単語の出現パターン

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 $\supseteq$ パターン 2 $\supseteq$ 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．

3. 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．

3.1 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 2 つ（タスク 1, 2）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

*2 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/

ク 2 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 3）である．タスク 3 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．

3.1.1 タスク 1：小説の読書とネタバレの記述

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）．

3.1.2 タスク 2：他人の書いた文へのネタバレ度合いづけ

タスク 2 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．

3.1.3 タスク 3：ネタバレを構成するのに必要な文節の選択

タスク 3 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 2 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．

3.2 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．

3.2.1 タスク 1

タスク 1 によって得たネタバレの文の数を表 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．

3.2.2 タスク 2

タスク 2 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 3 に示している（表 3 中の “対象となる文数”）．

3.2.3 タスク 3

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 3 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

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

表 3: 4 人以上が 4 以上のネタバレ度合いをつけた文の数と，その文から抽出された単語の数（抽出単語数）

	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)

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

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 [4, 5, 6] 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．

4. 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．

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 4 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．

表 4，表 5 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全単語でパターン 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．

表 4: 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 5: 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

5. ネタバレ検出手法についての考察と今後の評価実験枠組み

この章では，まず実際のレビュー文書からネタバレを検出する手法について考察する．最後に，今後行うレビュー文書からのネタバレ抽出の実験についての枠組みを述べる．

5.1 調査の結果を基にしたネタバレ検出手法

調査の結果から，ストーリー文書の後半に偏って出現する単語（我々が定義した出現パターンに該当する単語）は，ネタバレの記述に使用される傾向があることが示唆された．そこで検出手法として，レビューの文章（文書単位，文単位で検出するかによって異なる）中のパターンに該当する単語の数（あるいは割合）に対して，閾値をもけて判別する手法が考えられる．例として,「モルグ街の殺人（item5）」のレビュー文書（下記「レビュー文書例」参照）に対して，パターンに該当する単語がどれほど含まれているか調べる．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．

レビュー文書例（「モルグ街の殺人」）
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..

「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．レビュー文書例で示したように，パターンに該当する単語は必ずしも重要な単語ではない．また，1 文に対してパターンに該当する単語の数も少ない．そのため，提案した手法では文単位でネタバレが含まれているかどうかを判定するのは困難だと考えられる．

5.2 今後の評価実験の枠組み

この節では，今後行うレビュー文書からのネタバレ抽出の実験について述べる．評価に使用するレビュー文書は Amazom.co.jp と Booklog，読書メータの 3 つの Web サイトから収集する．事前に，複数の評価者に，我々が指定した複数の小説を読んでもらい，各小説に対するレビューを自由に記述してもらう．この評価者が記述したレビューは，評価者が小説を丁寧に読んだかどうかを判断するために使用する．また，小説は青空文庫に掲載されている小説を利用し，上記の 3 つの Web サイトから収集できるレビュー文書の合計の数が大きい小説を選ぶ．

実験では，評価者に各小説のレビュー文書を 50 ずつ見せて，それぞれのレビュー文書内にネタバレが含まれているか否かを評価してもらう．どのような内容をネタバレとして評価するかは被験者に指示せず，被験者それぞれの主観に基づいて評価してもらう．また，ネタバレが含まれていると評価されたレビュー文書に対しては，該当箇所（ネタバレを含む文）の抽出も行ってもらう．評価者の多数決によってネタバレが含まれるレビュー文書の正解データを作成し，提案手法による精度や再現率の評価を行う予定である．

6. 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傾向が見られた．今後は，実際にユーザがレビューを見た時に，提案手法がどれだけ有効かを定量的に分析するための実験を行う．

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

[5] 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).

[6] 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).