ストーリー文書内のネタバレの記述に関する基礎的調査

前田 恭佑† 土方 嘉徳† 中村 聡史††

† 大阪大学大学院基礎工学研究科 †† 明治大学/JST CREST

k.maeda@nishilab.sys.es.osaka-u.ac.jp hijikata@sys.es.osaka-u.ac.jp satoshi@snakamura.org

概要 Amazon.com のようなショッピングサイトでは，商品やコンテンツ（アイテム）に対して，感想や意見（レビュー）を閲覧，作成することができる．小説や映画などのストーリーのあるアイテムに対してのレビューには，そのストーリーの内容に関する記述が含まれていることがある．その中には，アイテムをまだ購入していないユーザの楽しみを奪ってしまうような記述（ネタバレ）が存在する可能性がある．我々は，ネタバレがストーリーの展開における位置づけや役割と関係があると考えた．そこで，レビュー文書とは別に，アイテムのストーリーの展開がわかる文書を利用する．本研究では，アイテムごとのストーリーを利用したネタバレの検出手法を提案する．

キーワード ネタバレ検出，ユーザレビュー，評判情報分析，計量言語学

1 はじめに

近年，一般ユーザがある商品やコンテンツ（以降，アイテム）に対して，意見や感想（以降，レビュー）を Web 上で公開し他のユーザと共有することが盛んになりつつある．一般 l 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有益な情報となりうる．しかし，コミック，小説，映画などのストーリー性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，その作品の結末や詳細なストーリーの展開，犯人の名前を挙げるなど，実際にアイテムを体験した時の感動や楽しみを減らしうる記述が存在する．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 [1] 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 [2] を検出する研究が行われている．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文には，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は有益な記述といえる．

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展開における位置づけや役割により大きく異なる．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（できれば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展開に関する情報が含まれていないため，各記述がユーザに与える影響を予測することは困難であった．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（以降，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．ストーリー文書があれば，アイテムの内容に関する記述が，ストーリー展開上重要なものなのか，そうでないかを判定することができる可能性がある．

しかし，ストーリーの展開において，どのような内容に関する記述がユーザに悪影響を与えるか（ネタバレとして認識されるか）は分かっていない．そこで本研究ではこの調査のために，いくつかのアイテムに対して，複数人の評価者によりそれを閲覧・視聴してもらい，ネタバレとなる内容について記載してもらうことにした（これを正解データとする）．我々は，ストーリー文書中の位置がネタバレと関連があると仮定して，正解データ中の記述がストーリー文書中にどのような分布で出現するかを調査した．

本稿では，2 章でデータセットの作成方法について述べる．3 章で調査手法について述べる．4 章でその結果と考察について述べる．最後に 5 章でまとめを述べる．

2 データセット

2.1 対象としたアイテム

ストーリー性を持つアイテムは映画，小説，コミックなどさまざまな種類が存在する．その中で，我々は青空文庫に掲載される小説を対象とした．理由は，アイテム

[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

のストーリー文書として小説の本文を利用できるためである．本研究では，青空文庫分の分野別リスト中の小説・物語カテゴリに属するアイテムから 5 つを選んでいる．

2.2 正解データ

6 人の学生を評価者として，青空文庫から指定した小説 5 つを読んでもらう．そして，アイテムごとにそれぞれのユーザが考えるネタバレとなる文（以降，ネタバレ文）を箇条書きで可能な限り書いてもらう．ネタバレは，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指示した．次に，ユーザがネタバレ文を書く際に共通に利用した単語に注目する．評価者間で共通して出現する回数が過半数である 4 回以上の単語を，そのアイテムにおいてネタバレと強く関連する単語とした．これらの単語を集めたものをネタバレの正解データとした．

3 調査手法

我々は，ネタバレに関連する内容がストーリー文書の後半部分に偏って出現すると仮定した．そこで，ストーリー文書を先頭から末尾の方向に複数個に分割し，各部分での単語の出現頻度を求めた．ストーリー文書全体で均等に出現する（各分割部分に等しく出現する）単語を BaseLine として，各単語をこれと比較することで分布の偏りを調べた．図 1 は単語の出現割合の累積をグラフにしたものである．図 1 の単語 1 のように，BaseLine の下にあれば後半に偏って出現する単語とする．単語 2 のように，BaseLine よりも上にあれば前半に偏って出現する単語とする．そして，正解データの中に後半に偏って出現する単語がどれほど存在するか，その単語はどのようなものなのかを調べる．

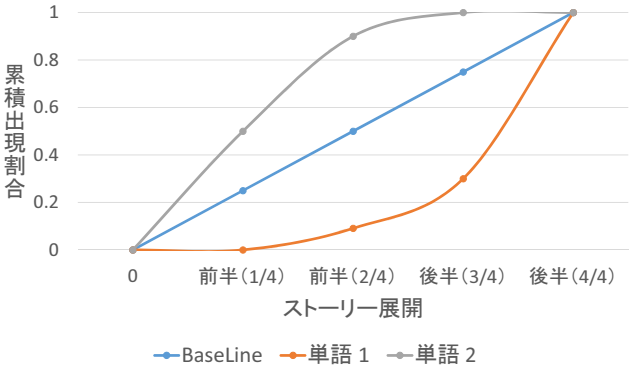

図 1 単語の出現の仕方

4 結果・考察

はじめに正解データ中の単語がどのようなものであったかを定性的に説明する．まず，人物名や地名などの，他のアイテムでは出現しないアイテムに特有の単語がいくつか存在した．従来研究でもアイテムに特有の単語はストーリーの内容に関する記述で利用されやすいと示唆されている [2]．しかし，アイテムに特有の単語ではない単語も多くあった．また，ストーリー文書で用いられない単語も正解データとして入っていた．これらはストーリー文書中の表現の言い換えや，ストーリー文書に書かれていないことを評価者が推測して書いた言葉であった．現在はストーリー文書中で使われない表現には対応することができていない．この解決は今後の課題となる．

調査の結果，正解データの中で後半に偏って出現する単語の割合は平均で 0.45 となった．これは，ネタバレ文が後半に偏った単語のみで構成されているわけではなく，後半に偏った単語とそうでない単語の組み合わせによって構成されている可能性があるといえる．たとえば，登場人物の行動について書かれていた場合，行動内容は後半に偏った単語が使われているが，登場人物は後半に偏った単語ではないということが考えられる．実際に，後半に偏っていなかった単語は，主人公などのストーリー序盤から均等に登場する人物名や，小説での舞台を示す単語であった．これは正解データ作成時に，係り受けまでみて単語を考慮すべきといえる．

5 おわりに

本研究では，ネタバレの検出にストーリー文書を利用することを検討した．実験においては，信頼性や普遍性の観点からストーリー文書として小説を利用した．人物名や地名などのアイテム独自の単語が正解データに含まれることは従来研究に一致するものであった．また，ネタバレ文が後半に偏った単語のみで作られるわけではないことが示された．

今後は，別の方法でネタバレ文から正解データを作成することを考えている．まず，評価者に他の評価者が書いたネタバレ文を提示し，1 から 5 の 5 段階でネタバレの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）．ここでネタバレ度合いの平均値が高いネタバレ文は，大多数のユーザが考えるネタバレと言える．こうして得られた文の単語に着目することで，より妥当な正解データが作成できると考える．

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] Guo, S. and Ramakrishnan, N.: Finding the story-teller: automatic spoiler tagging usin linguisitic cues. Proc. of Coling'10, pp. 412–420, 2010.

[2] 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.