

推薦論文

レビューの文脈一貫性を用いたあらすじ文判定手法

岩井 秀成¹ 土方 嘉徳^{1,a)} 西田 正吾¹

受付日 2013年12月22日, 採録日 2014年4月1日

概要: Amazon.com に代表されるショッピングサイトの多くでは、アイテムに対してレビューを簡単に作成・閲覧できる機能が提供されている。レビューに書かれた他者の意見は有用であるが、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、ストーリーの内容（本稿では「あらすじ」と呼ぶ）が書かれている場合がある。レビューによりあらすじが分かると、実際に小説や映画を見たときの楽しみや感動が減ってしまい問題である。そこで我々は、あらすじを文単位で検出し除去することを試みる。我々は、このあらすじ検出を実現するために、各文がレビューとしての文脈を維持する形で書かれていること（文脈一貫性）に注目する。本研究では、レビュー文書の文書構造とあらすじ以外の観点に基づくクラス情報を用いたあらすじ文判定手法を提案する。評価実験を行い、文書中における文の位置情報の付与と文脈情報の利用によりあらすじ判定の精度が向上することを確認した。また、被験者実験により、提案手法はユーザが知りたくなかった記述の多くを除去できており、かつ、商品購入の際に意思決定につながる文の多くを表示できていることを明らかにした。

キーワード: 評判情報分析, あらすじ判定手法, 機械学習, ユーザレビュー, 文脈一貫性

Plot Classification Method Using Contextual Coherence for Review Sentences

HIDENARI IWAI¹ YOSHINORI HIJIKATA^{1,a)} SHOGO NISHIDA¹

Received: December 22, 2013, Accepted: April 1, 2014

Abstract: Users can write and read reviews of items easily on many Internet shopping sites. Reviews on items like novels and movies can include spoilers (undesired plot descriptions) along with the opinions of the review authors. It is an issue that spoilers might reduce a delight in reading novels and watching movies. Therefore, we try to detect and eliminate spoilers by sentence. In this paper, our target is utilization of contextual coherence that is originally embedded in a review document and makes readers to understand the content easier. We propose a plot classification method using the document structure and the information of multiple classes in reviews. Experimental results show that using position information of a sentence and context information contributes to the improvement of classification results. In addition, a user study shows the proposed method can remove spoilers without removing useful sentences.

Keywords: sentiment analysis, plot classification, machine learning, user reviews, contextual coherence

1. はじめに

Amazon.com に代表されるショッピングサイトの多くでは、アイテムに対してレビューを簡単に作成・閲覧できる機能が提供されている。レビューに書かれた他者の意見は

有用であるが、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、ストーリーの内容（本稿では「あらすじ」と呼ぶ）が書かれている場合がある。レビューによりあらすじが分かると、実際に小説や映画を見たときの楽しみや感動が減ってしまい、問題である。この問題を解決するには、あらすじを検出し、それを除去す

¹ 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University,
Toyonaka, Osaka 560-8531, Japan

a) hijikata@sys.es.osaka-u.ac.jp

本稿の内容は 2013 年 11 月の WebDB フォーラム 2013 にて発表され、同シンポジウムプログラム委員会により情報処理学会論文誌データベースへの掲載が推薦された論文である。

る必要があると考える。先行研究 [1] では、ストーリーを持ったアイテムに対して、レビュー文を対象としたあらすじ判定手法が提案されている。その手法では、ある文があらすじか否かを判定するために、その文に含まれる単語のあらすじらしさのみを用いた。しかし、その文が属するレビュー文書中での、その文の周辺情報も用いると、その判定の精度が向上すると思われる。我々が特に注目したのは、各文がそのレビューが持つ文脈を維持する形で書かれている（文脈一貫性がある）と思われる点である。具体的には、その文があらすじか否かは、その文に含まれるレビュー文書の構造に関する情報や、あらすじ以外の観点でその文に付与（推定）されたクラス情報とも関係があると思われる。

本研究では、前者の情報としてレビュー文書の文書構造に注目した。文書構造とは、文書の長さや段落の数、各文の文書中での位置など、レビューの内容にはよらない文書の表層的情報のことである。本研究では文書構造として、判定したい文（注目文）が含まれるレビュー文書中の文の数（文数情報）、注目文のレビュー文書中での位置（位置情報）、注目文の前後の文のあらすじらしさ（文脈情報）を取り上げる。また、後者の情報として、評判情報分析でよく用いられる意見か否かを表すクラス（以降、意見クラス）に注目した。評判情報分析では意見か否かはレビューの重要な情報と見なされており、あらすじ判定の有用な手がかりになる可能性がある。これらの情報とあらすじとの関係を調査することで、あらすじ判定に有用な特徴を発見し、それらを用いて判定の精度向上を図る。

本稿で、あらすじはストーリーに関する記述とする。ストーリーに関する記述が書かれていれば、それが概要レベルのものであるか詳細な内容であるかにかかわらず、また、記述量が多いか少ないかにかかわらず、すべてをあらすじと見なす。また、本研究では英語のレビューを対象とする。

本稿の章構成は以下のとおりである。2章で関連研究を紹介する。3章で先行研究 [1] のあらすじ判定の基本手法を説明する。また、レビュー文書の文書構造と意見クラスの利用があらすじ判定に貢献する可能性を説明する。4章で本研究に用いるデータセットについて説明する。5章でレビュー文書の文書構造とあらすじとの関係、意見クラスとあらすじとの関係を調査する。6章で提案手法について説明する。7章でその提案手法の評価実験を行う。最後に8章でまとめを述べる。

2. 関連研究

先行研究 [1] では、ストーリーを持ったアイテムに対するレビュー文を対象としたあらすじ判定手法が提案された。また、その手法を用いてあらすじを非表示にするシステムが実装された。被験者実験により、このシステムが spoiler（ストーリーに関する重要な記述でユーザが見たくないもの）の遮蔽に有効であることが検証された [2]。その先行研

究では、文を bag-of-words で表現し、機械学習により判定を行っている。つまり、単語のあらすじらしさのみから各文のあらすじらしさを求めており、文書全体の構造や他のクラスに関する情報は用いていない。本研究の貢献は、レビュー文書の文書構造や他のクラスに関する情報を、あらすじ判定に初めて適用したことにある。本章では、あらすじ判定に関連する研究として、意見判定に関する研究と、レビュー文書の文書構造に着目した研究を取り上げ、紹介する。

2.1 意見判定に関する研究

インターネット上のレビュー文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。中でも、我々の研究に最も関連する分野は文書内の意見に着目した研究分野である。Pang ら [3] は、レビュー中の各文に対して Naive Bayes, SVM を用いて主観的な文か客観的な文かを判定することで、どの箇所に意見が書かれているか特定している。Yu ら [4] は、上記の判定を文単位と文書単位の両方で行えるようにしている。Riloff ら [5] は、上記判定のためにラベル付けされていないデータから学習するブーストラップ手法を提案している。

レビューが肯定的な文か否定的な文かを判定する研究もある。Dave ら [6] は、Naive Bayes を用いてレビュー文が肯定的か否定的かを判定している。Wilson ら [7] は、単語と文脈の両方の評価極性を考慮し、句ごとに判定している。

本研究では、意見判定の推定結果（以降、意見らしさのスコア）を用いて、あらすじ判定の精度向上を目指している（6.2 節参照）。その意見らしさのスコアの算出には、Naive Bayes または SVM による教師あり学習手法を用いているが、意見らしさのスコアが算出できれば、この節で述べた他の意見判定手法を用いてもかまわない。

2.2 文書構造に着目した研究

意見判定および一般的な文書分類においては、文書中の文の位置に注目することで判定（分類）精度が向上することが分かっている。Taboada ら [8] は、文書中において、書き手の主要な意見は文書全体に均等に現れるのではなく、特定の部分に集中して現れるという知見を示した。意見に対する評価表現が出現する位置に応じて、各評価表現が持つ評価極性値を修正することで、意見判定の精度を向上させた。江崎ら [9] は、Weblog 記事中の賛否に対する評価表現が出現する位置に応じて、評価極性値を修正することで、賛否判定の精度を向上させた。Otsubo ら [10] は、文書構造の上位階層に位置するテキストと、アンカ周辺の文章を利用することで、Web ページのカテゴリ分類の精度を向上させた。

また、判定対象の文だけでなく、前後の文に注目することでも判定精度が向上することが分かっている。那須川

ら [11] は、逆接関係を導く表現が存在しなければ、ある文の意見についての評価表現の評価極性とその周囲の文の評価表現の評価極性は等しい傾向があるという知見を示した。梅澤ら [12] は、Naive Bayes によりクレーム文の検出を行った。評価表現に基づいてクレームとして検出された文の前後に位置する文も、同様にクレームであると考え、これを学習データに加えることでクレーム文の検出性能が向上することを示した。

これらの研究では、文書構造に関する情報が判定や検出の精度向上に貢献している。このことに着想を得て、あらすじの判定においても文書構造に関する情報が有効に働く可能性があると考えた。そこで本研究では、単語の情報に加えて、文書構造に関する情報も用いてあらすじ判定を行い、その精度向上を目指す。

3. 機械学習によるあらすじ判定

本章では、先行研究 [1] でのあらすじ判定を行うための基本手法を紹介する。また、レビュー文書の文書構造と意見クラスの利用があらすじ判定に貢献する可能性を説明する。

3.1 あらすじ判定の基本手法

先行研究 [1] では、レビュー文書中の単語を素性として機械学習を行うことで、ある文があらすじか否かを判定する手法が提案されている。この手法では、文にあらすじか否かを表すクラスが付与された学習データを用いて判定モデル（あらすじ分類器）を学習する。そして、学習されたモデルを用いて未知の文にあらすじらしさのスコア付けを行い、あらすじか否かを判定する。また、機械学習を行うために、文 p を bag-of-words で表現した。すなわち、文 p は、単語 w_m により、 $p = \langle w_1, w_2, \dots, w_M \rangle$ のように表現される。また、文 p_n 中で各単語 w_m が出現した回数 $x_{n,m}$ も記録した。

機械学習アルゴリズムには、学習されたモデルによってあらすじらしさを示すスコアを得やすいものを使用した。先行研究 [1] は、Naive Bayes, SVM, Logistic 回帰, 決定木, k-nearest neighbor 法の 5 つのアルゴリズムを用いた。評価実験により、あらすじ判定には Naive Bayes と SVM が有効であることが分かった。本研究で提案するあらすじ判定手法においても、先行研究の手法をベースとする。また、先行研究での評価実験の結果から Naive Bayes [13] と SVM [14] の 2 つを用いる。

先行研究 [1] では、レビュー文書中の文章を文に分割している。本研究で提案するあらすじ判定手法でも、文章を文に分割し、文があらすじか否かを判定している。本研究は、先行研究 [1] の文への分割手法と同じものを用いている。ここでは、その分割手法について説明する。基本は、レビューにおいて「.」「!」「?」が出現し、その後スペース

もしくは改行が存在する場合に分割を行う。ただし、「.」「.」のようにピリオドとスペースもしくは改行の間に右括弧、クォーテーション、ダブルクォーテーションが挟まる場合があるため、これらが存在した場合にも分割を行う。また、「i.e.」「P.S.」「e.g.」「etc.」のような特定の省略語にはピリオドとスペースの組合せが登場するため、これらの省略語では分割を行わない。本研究でのあらすじ判定手法では単語のSTEMMINGを行うが、STEMMINGをして単語が残らない文（たとえば、特殊記号のみが含まれる文）は、データから除外される。

3.2 レビュー文書の文書構造と意見クラスの利用

我々は、あらすじ判定に単語のあらすじらしさだけでなく、レビュー文書の文書構造と意見クラスを用いることで、判定精度が向上するのではないかと考えた。具体例を用いてこの仮説を説明する。以下にレビュー文書の例を示す。この文書中の“*She has created a wonderful world filled with the most charming, fascinating people.*”をあらすじか否かを判別したい注目文とする。

(レビュー文書の例) I learned about Tasha Alexander's book from a list I'm on, and went to her website to read the first chapter. She has created a wonderful world filled with the most charming, fascinating people. It was so delightful, I decided I'd splurge and buy it. This is one book I know I'm going to re-read. Some of the other reviews say she is going to write a sequel, I sure hope they're right.

先行研究 [1] の評価実験では、あらすじ判定の精度を評価するとともに、どのような単語があらすじ判定に貢献するかについても調査されている。その結果、人名とその代名詞（すなわち“he”や“she”）が出現する文は高い確率であらすじであることが分かった。これは、レビューがあらすじについて記述するときには、「誰がどうした」という形式で記述するためである。この注目文には“she”という単語が含まれるため、あらすじと判定される可能性が高いが、実際にはあらすじではないため、これは誤判定となってしまう。

しかし、注目文が含まれるレビュー文書の文数情報、注目文の位置情報と文脈情報を用いることで、あらすじである確率を低く推定できると思われる。まず文数情報であるが、レビュー文書中の文の数が少ないと、その中に含まれる文にはあらすじが記載されている確率が低くなる可能性がある。これは、レビューがレビューを書く際に、まずは自身の感想を述べることを優先するためである。そのため、文の数が少ないとそのほとんどが感想を述べるために用いられると考えられる。この例では、レビュー文書中の

文の数は5つと少ないので、その文書中の各文があらすじである確率は低いと思われる。

また、位置情報であるが、レビュー文書中での位置がレビュー文書の先頭と末尾に近いほどあらすじでないと思われる。これは、レビューの冒頭にあらすじが少ないのは、先に述べたように、まずはレビューの感想を述べるのが優先されるためである。また、レビューの末尾にあらすじが少ないのは、もう1度レビューの感想を述べて、レビューのまとめにするためである。この注目文は先頭から2番目に書かれているので、あらすじである確率は低いと思われる。

最後に、文脈情報であるが、あらすじはまとめて書かれやすい可能性がある。前後の文があらすじである場合、挟まれた文もあらすじらしさが高いと考えられる。これは、急な文脈の変化が頻繁に起こるような文章は読みにくく、人は内容のつながりを重視して文章を書くからである。この注目文の前後の文はあらすじではないので、注目文があらすじである確率は低いと思われる。

また、意見クラスに関する情報に注目すると、あらすじである確率はより低くなると思われる。それは、意見とあらすじは同じ文に登場しにくいと考えられるからである。評判情報分析の分野では、文や文書が意見を述べているか否か(客観的内容か主観的内容か)に判定することがよく行われている。そこで、その意見判定の結果を利用すれば、誤判定を防げる可能性がある。この注目文は、“wonderful”, “charming”, “fascinating”という、意見に用いられやすい単語を含んでいる。そのため、この文は意見と判定される確率が高くなり、結果としてあらすじである確率は低いと思われる。

これらの仮説が正しいかどうかを5章の調査により明らかにする。

4. データセット

本章では、文書構造と意見クラスに関する調査と提案手法の評価実験を行うのに用いる、あらすじ(および意見)に関するデータセットについて述べる。我々は、評価者にレビュー内の各文があらすじか否かの判断を行ってもらい、あらすじに対する正解データを作成した。正解データの作成方法は以下のとおりである。我々はまず、ストーリーを持ったアイテムを持つカテゴリとしてコミック、小説、DVDを選択した。そして、Amazon.comから各カテゴリに属するアイテムをランダムに100個ずつ選択した。選択された各アイテムに付けられたレビューからランダムに5件ずつ選択した。このようにして各カテゴリからそれぞれ500件のレビューを取得した。その後、取得したレビューに対し、3人の評価者にラベル付けを行ってもらった。各評価者は、各文があらすじか否かを判定する。3人の評価者のうち2人以上があらすじであると判定した文をあらす

表1 各カテゴリのデータセットの詳細

Table 1 The details of the dataset in each category.

	コミック	小説	DVD
単語数	6,414	6,334	6,966
単語数(出現回数2回以上)	3,603	3,539	3,761
あらすじである文の数	1,523	1,602	1,357
あらすじでない文の数	3,484	3,225	3,445
意見である文の数	2,770	2,567	3,065
意見でない文の数	2,237	2,260	1,737

じと見なし、正解データを作成した。表1に、単語数、出現回数が1回のもので除いた単語数、あらすじである文の数と、あらすじでない文の数を示す。この表の単語数は種類数である。

なお、3人の評価者の判定の一致の程度を調べるためにFleissの κ 係数[15]を計算した。すると、コミックカテゴリで0.612、小説カテゴリで0.544、DVDカテゴリで0.466となった。一般に、 κ 係数は、0-0.4で低い一致、0.4-0.6で中程度の一致、0.6-0.8でかなりの一致、0.8-1.0で高い一致といわれており[16]、今回のラベル付けでは中程度からかなりの一致ということが出来る。あらすじかどうかの判定において、3人の評価者が一致しない場合について調査した結果、その作品のメインテーマに関する情報や、主要な登場人物やキャラクターの立ち位置などの概要レベルのあらすじが多かった。これらは、その作品に興味のあるユーザであればすでに知っている可能性が高い。正解データを作成する際には多数決を行うことで、より妥当に判定されているであろうラベルを用いることにした。

また、評価者にレビュー内の各文が意見であるか否かの判断を行ってもらい、意見に対する正解データを作成した。ここで意見とは、レビューのアイテムに対する感想や要望とする。意見に対する正解データの作成方法は、あらすじに対する正解データを作成した場合と同様に、同じレビューデータに対して、3人の同じ評価者に各文が意見であるか否かのラベル付けを行ってもらった。3人の評価者のうち2人以上が意見であると判定した文を意見と見なし、意見に対する正解データを作成した。表1に意見である文の数、意見でない文の数を示す。Fleissの κ 係数を計算したところ、コミックカテゴリで0.531、小説カテゴリで0.551、DVDカテゴリで0.463となった。今回のラベル付けでは中程度の一致ということが出来る。

5. 文書構造と意見クラスの調査

本章では、レビュー文書の文書構造について、5.1節で文数情報と位置情報、5.2節で文脈情報を詳細に調査する。また、5.3節で意見クラスを詳細に調査する。これらの調査には、4章で作成したデータセットを用いている。

表 2 文の位置と文の数の例 (合計 5 文)

Table 2 An example of sentence location number (The number of sentences in the review is five).

位置	文の内容
1 文目	I learned about Tasha Alexander's ...
2 文目	She has created a wonderful world ...
3 文目	It was so delightful, I decided I'd ...
2 文目	This is one book I know I'm going to ...
1 文目	Some of the other reviews say she is ...

5.1 位置情報・文数情報の付与の調査

レビュー文の位置 (位置情報) とレビュー文書中の文の数 (文数情報) に注目し, あらすじとの関係性を調査する. 文の位置と文の数を各文に割り当てる. 表 2 に, ある文書における各文の位置と文の数の例を示す. ある文に注目したとき, レビューの上端もしくは下端の近い方から何文目かが「文の位置」である. また, レビュー全体の文の総数を「文数」とする. 同じレビュー文書中ではすべての文に対して同じ値が付与される. この例では, すべての文に対して, 文数として 5 が付与されている.

4 章のデータセット中のすべての文について, 位置情報と文数情報を調べ, 各位置に属する文ごとに, また各文数のレビュー文書に属する文ごとに, あらすじである文の割合を調査した.

図 1 にレビュー文書の文の数ごとの文の位置とあらすじである文の割合を示す. 折れ線グラフが, あらすじである文の割合を表している. この図から, 文の位置が大きいほど, あらすじを含みやすいことが分かる. また, レビュー文書の文の数にかかわらず, 文の位置ごとのあらすじである文の割合がほぼ同じ値をとることから, 文の位置があらすじである文の割合に強い影響力を持っていることが分かる. そして, レビュー文書の文の数が少ないほど文の位置が低いものしか含まないことから, レビュー文書の文の数が少ないほどあらすじである割合が少ないだろうという予想がつく. そのためレビューの文数があらすじである文の割合に弱い影響力を持っていることが分かる. しかし, この結果は文の位置による影響が表れた結果であり, 文数情報と位置情報を同時に用いることは有効でないと思われる. 文の位置とレビューの文数を独立に調査した結果を下に示す.

図 2 に文の位置とあらすじである文の割合の関係を示す. レビューの上端に近い位置とレビューの下端に近い位置を区別することにする. 文数が奇数でありかつ中央の文はどちらにも近いとはいえないため除外して算出している. この図では, 各位置に属する文の個数を棒グラフで示している. 折れ線グラフが, あらすじである文の割合を表している. 折れ線グラフが右肩上がりであることから, レビューの上端と下端から離れた場所に書かれているほど, あらすじを含みやすいことが分かる. また, 上端の折れ線

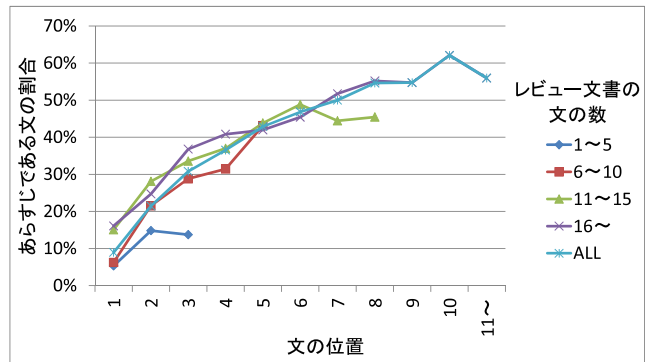


図 1 レビューの文の数ごとの文の位置とあらすじである文の割合
Fig. 1 The ratio of plot sentences in each sentence location.

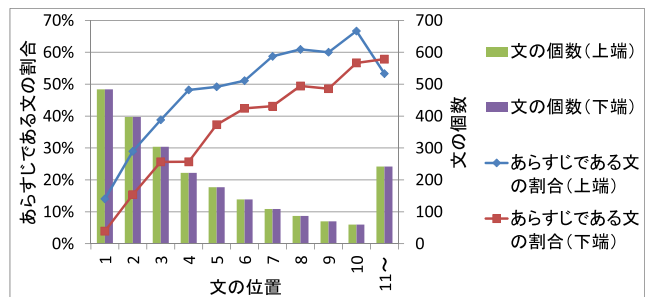


図 2 文の位置とあらすじである文の割合の関係
Fig. 2 The ratio of plot sentences in each sentence location when dividing the location to upper place and lower place.

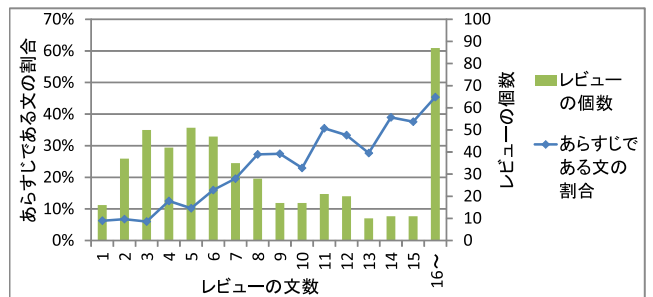


図 3 レビューの文数とあらすじである文の割合の関係
Fig. 3 The ratio of plot sentences according to the number of sentences in the review.

グラフ (青の線) の方が下端の折れ線グラフ (赤の線) よりも上にあることから, レビューの後半よりレビューの前半の方が, あらすじを含みやすいことも分かる.

図 3 にレビューの文数とあらすじである文の割合の関係を示す. 横軸が各レビュー文書中の文数である. 棒グラフは, 各文数のレビューの個数である. 折れ線グラフが, 各文数のレビュー文書に属する文のうち, あらすじである文の割合である. 折れ線グラフが右肩上がりであることから, 文数が多いほどあらすじを多く含むことが分かる.

以上より, 位置情報と文数情報は, その文があらすじであるか否かと関連があることが分かる. また, 位置情報と文数情報の間にも相関があるが, 位置情報の方があらすじ

表 3 文脈ごとのあらすじ割合

Table 3 The ratio of plot sentences in each context pattern.

前後1文 パターン	あらすじ 割合	出現 回数	前後2文 パターン	あらすじ 割合	出現 回数
TXT	90%	916	TTXTT	94%	591
			FTXTT	84%	115
			TTXTF	81%	162
			FTXTF	79%	48
TXF or FXT	50%	786	TTXFT	69%	64
			TFXTT	67%	54
			TFXTF	52%	42
			FTXFT	55%	38
			FFXTT	50%	145
			TTXFF	51%	219
			FFXTF	42%	106
			FTXFF	38%	118
FXF	9%	1,299	TFXFT	27%	30
			TFXFF	16%	161
			FFXFT	15%	110
			FFXFF	7%	998

との関連性が高いことが分かった。これらのことから、本研究で提案するあらすじ判定手法では、位置情報を利用する。機械学習で位置情報を付与する際には、文数が奇数でありかつ中央の文は便宜上、レビューの上端からの位置を算出し付与することにする。

5.2 文脈情報の利用の調査

注目する文の前後の文のあらすじらしさ（文脈情報）に注目し、注目する文のあらすじとの関係性を調査する。具体的には、注目する文 X の前後の1文または2文が、あらすじか否かを調べる。あらすじであった場合を T と表し、あらすじでなかった場合を F と表す。前後1文または2文のすべてのパターンごとに、注目文 X があらすじである割合を調べる。

表 3 に結果を示す。前後1文がともにあらすじであるとき、90%という高い割合で、注目文はあらすじである。逆に、前後1文がともにあらすじでない文のときは、注目文があらすじである割合はわずか9%である。このことから、注目文の前後1文は、注目文があらすじか否かに強く影響していることが分かる。また、前後2つ目の文があらすじか否かも、前後1つ目の文に比べて弱いものの、注目文があらすじか否かに影響している。

以上より、前後2文（前後1文目と前後2文目の両方）に対してあらすじ判定を行ったときに得られたあらすじらしさのスコアを利用することで、注目文のあらすじ判定の精度が向上すると考えられる。

5.3 クラスの割合に関する調査

意見とあらすじが同じ文に登場しにくいかどうか調査す

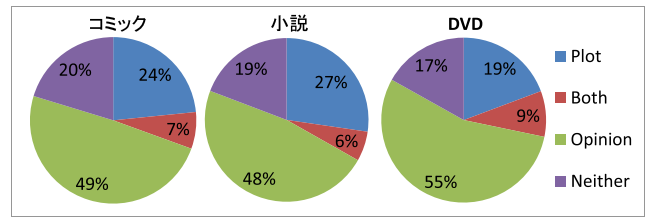


図 4 あらすじと意見の関係

Fig. 4 The relationship among plot and opinion in one sentence.

る。我々のデータセットには、各文にあらすじか否かと、意見か否かに関するラベルが付いている。データセット中のすべての文に対して、あらすじラベルのみ付いているもの (Plot)、意見ラベルのみ付いているもの (Opinion)、両方のラベルが付いているもの (Both)、どちらのラベルも付いていないもの (Neither) の数を調べ、データセット中のすべての文に対する割合を求める。

なお、両方のラベルが付いているものの例をあげると、“I could never forget that they had known each other a matter of days when they were overcome with love.”がある。この文では、忘れられないという意見とその内容の両方が書かれている。また、どちらのラベルも付いていない例をあげると、“The author of this book was born in Florida.”がある。作者がフロリダ州出身であるという客観的事実が書かれている。

図 4 に、あらすじと意見の関係を示す。Both の割合はどのカテゴリにおいても10%以下である。このことから、レビューにおいて意見とあらすじは異なる文に書かれることが多いことが分かる。よって、意見らしさを推定し、そのスコアを利用することで、あらすじ判定の精度が向上する可能性がある。

6. 提案手法

6.1 ベースラインの詳細

本研究では、5章の調査結果に基づき、先行研究 [1] をベースとして、レビュー文書の文書構造と意見クラスの情報を用いたあらすじ判定手法を提案する。まず、基本となる先行研究の手法（以降ベースライン (Baseline)）の詳細を述べる。3.1節で先行研究の基本的な判定手法を示したが、実際にはいくつかの工夫を行っていたのでそれについて説明する。この工夫を行ったものをベースラインとする。

図 5-(a) にベースラインにおいて、あらすじ分類器を学習し、新しいレビュー文をあらすじか否かに判定する過程を示す。本研究では、このあらすじ分類器（後に、文脈情報を利用する場合と意見判定を利用する場合で用いる分類器）を「基本あらすじ分類器 (A)」と呼ぶ。

ベースラインでは、bag-of-words で表現された文とそのあらすじに関するラベルを用いて、基本あらすじ分類器

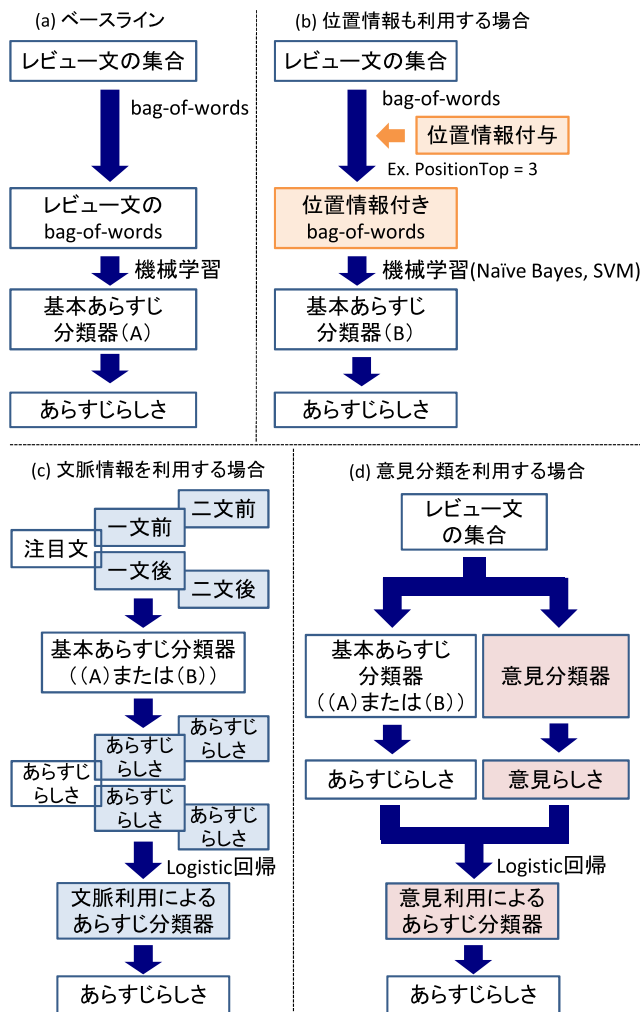


図 5 ベースラインと提案手法の処理の流れ

Fig. 5 The process flows of the baseline and the proposed method.

(A) を学習する。そして、この基本あらすじ分類器 (A) を用いて、未知の文にあらすじらしさを表すスコアを与える。機械学習のアルゴリズムは、先行研究 [1] で高い判定性能を示した Naive Bayes [13] と SVM [14] を用いた。

また、先行研究 [1] の評価実験で、人名（登場人物名、著者名、俳優名とその他人名）と特有語（ある特定のアイテムにしか出現しない特有の単語）の一般化があらすじ判定に有効であることが分かっている。人名や特有語は、固有名詞であることが多いため、多くのアイテムのレビュー文書に出現しない。そのため、出現した人名や特有語をそのまま素性として用いても、他の文書中の文の判定に貢献しない。そこで、自然言語処理により、これらを <character> や <peculiar> のようなタグで一般化することにより、多くの文で有効な素性になるようにした。

人名や特有語の一般化の詳細は先行研究 [1] に示されている。ここでは手法の概要とその精度について簡単に紹介する。人名の一般化については、米国国勢調査局の人名データベースから一般語を除いたものを人名と見なした。

表 4 人名抽出の精度、再現率

Table 4 The precision and recall of the human name extraction.

	コミック	小説	DVD
精度	0.945	0.967	1.000
再現率	0.520	0.492	0.786

また、Amazon.com 中のアイテムの紹介文に出現した人名を登場人物名とした。著者名と俳優名は Amazon.com のアイテムの紹介ページ中に決まった形式で書かれているため、テンプレートマッチングで抽出した。特有語の一般化については、複数のアイテムに出てこない単語を特有語を表すタグで置き換えることで一般化を行った。どの一般化においても定義どおりに抽出すること自体には問題がないが、上記の人名抽出には未検出が起こりうる。なぜならば、人名データベースに載っている人名でも、それが一般語としても用いられるものであれば、人名として取り扱わないからである。そこで、人名抽出に対する精度と再現率を表 4 に示す。どのカテゴリにおいても 100% 近い精度が得られているが、再現率はそれよりは低くなっている。

6.2 提案手法の詳細

提案手法はベースラインに、位置情報、文脈情報、および意見クラスに関する情報を用いて、分類器を強化したものである。評価実験では、上記の 4 種類の情報がそれぞれ、どれほど判定精度の向上に寄与するかを調べる。そのため、それぞれを個別に用いた場合での実装も行う。

図 5-(b) に位置情報を用いる場合の、あらすじ判定の過程を示す。提案手法では、判定したい文をその単語情報と位置情報を用いて、あらすじらしさを判定するためのモデルを学習する。このモデルを基本あらすじ分類器 (B) と呼ぶ（この分類器も、以降で文脈情報を利用する場合と意見判定を利用する場合で用いる）。手法の詳細は以下のである。文を bag-of-words で表現し、さらに位置情報を新たな要素として追加する。たとえば、注目文がレビューの上端から 3 文目の場合、*PositionTop* = 3 という素性が付与される。文の位置は大きくなるほど存在数が減少していくため binning を行う。位置 11 以上は位置 11 と見なす。その後、Naive Bayes または SVM により機械学習を行う。正解データは、あらすじに関するラベルである。これにより基本あらすじ分類器 (B) が構築される。

図 5-(c) に文脈情報を用いた場合の、あらすじ判定の過程を示す。注目文とその前後 2 文を、基本あらすじ分類器 (A) または (B) にかけて、あらすじらしさのスコアを求める。単語情報に加え文脈情報のみ用いる場合は、基本あらすじ分類器 (A) を用いる。単語情報に加え、位置情報と文脈情報を用いる場合は、基本あらすじ分類器 (B) を用いる。これで 5 つのあらすじらしさのスコアが得られるが、これら

を「文脈利用によるあらすじ分類器」にかけて、最終的なあらすじらしさのスコアを求める。ここで、文脈利用によるあらすじ分類器の構築方法であるが、これは事前に判定して得られた5つのあらすじらしさのスコアと注目文のあらすじに関する正解ラベルから Logistic 回帰により学習する。このデータは、基本あらすじ分類器を構築するのに用いた学習データとは別に用意しておく（データセット分割の詳細は 7.1 節で説明する）。

図 5-(d) に意見情報を用いた場合のあらすじ判定の過程を示す。判定したい文を、基本あらすじ分類器 (A) または (B) にかけて、あらすじらしさのスコアを求める。単語情報に加え意見情報のみ用いる場合は、基本あらすじ分類器 (A) を用いる。単語情報に加え、位置情報と意見情報を用いる場合は、基本あらすじ分類器 (B) を用いる。また、意見分類器にもかけ、意見らしさのスコアも求める。ここで、意見分類器とは、文を意見か否かに判別するモデルである。この分類器の構築方法であるが、文を bag-of-words で表現し、学習データセットの意見に関するラベルを用いて機械学習する。ただし、位置情報は埋め込まない。機械学習のアルゴリズムは、Naive Bayes または SVM を用いる。

あらすじらしさと意見らしさの2つのスコアを得るが、これを「意見利用によるあらすじ分類器」にかけて、最終的なあらすじらしさのスコアを求める。ここで、意見利用によるあらすじ分類器の構築方法であるが、これは事前に判定して得られたあらすじらしさのスコアと意見らしさのスコアとその文のあらすじに関する正解ラベルから Logistic 回帰により学習する。このデータは、基本あらすじ分類器を構築するのに用いた学習データとは別に用意しておく（データセット分割の詳細は 7.1 節で説明する）。ここで意見らしさのスコアとあらすじらしさのスコアは、同一のアルゴリズムから算出されたものを使用する。たとえば、あらすじらしさのスコア算出に SVM を使用する場合は、意見らしさのスコア算出にも SVM を使用する。

最後に、文脈情報と意見クラスの情報を両方とも利用する場合の実装方法について述べる。なお、7章で行う実験では、意見クラスの情報を利用してもあらすじ判定の性能が向上しなかったため、この組合せの実験は行っていない。文脈情報を利用すると、基本あらすじ分類器 ((A) または (B)) により、注目文と前後2文のあらすじらしさのスコアが得られる。また、意見判定を利用すると、注目文の意見らしさのスコアが得られる。これら6つのスコアとあらすじに関する正解ラベルから Logistic 回帰により、最終的な分類器を学習する。

7. 提案手法の評価

本章では、6章で説明した提案手法の評価を行う。位置情報、文脈情報、意見判定を利用した場合としない場合、またそれらの組合せについて、どれだけの効果があるのか

を調べる。

7.1 学習条件と評価尺度

10分割交差検定法 (ten-fold cross validation) [17] で評価を行った。ただし、提案手法では利用する分類器が階層的に存在するため、データを3つに分割する変則的な交差検定を行うことにした。基本あらすじ分類器と意見分類器を構築するのに使うための learning set (10分割中の8つを用いる)、文脈利用によるあらすじ分類器と意見利用によるあらすじ分類器を構築するのに使うための validation set (10分割中の1つを用いる)、最後に評価を行うための test set (10分割中の1つを用いる) である。10分割交差検定におけるデータの分割はアイテム単位で行う。これは、同一のレビュー、同一のアイテムへのレビューが学習データと評価データの両方に同時に出現することを防ぐためである。

交差検定の方法について、より詳細に説明する。まず、learning set とあらすじの正解ラベルを学習データとして基本あらすじ分類器を、learning set と意見の正解ラベルを学習データとして意見分類器を構築する。次に、その基本あらすじ分類器と意見分類器を用いて、validation set と test set にあらすじスコアと意見スコアを付与する。その後、あらすじスコアと意見スコアを付与された validation set とあらすじの正解ラベルを学習データとして文脈利用によるあらすじ分類器と意見利用によるあらすじ分類器を構築する。最後に、文脈利用によるあらすじ分類器、もしくは意見利用によるあらすじ分類器を用いて、test set に最終的なあらすじスコアを付与する。test set の最終的なあらすじスコアによる判定があらすじの正解ラベルと一致しているか否かで分類器の性能を評価する。

一般に識別問題において、不均衡データは予測精度に悪影響を与えることがある。表 1 に示したように、得られた学習データにおいては、あらすじのラベルが付いた文が少ない。このようなデータへの対策には、以下の3つが知られている [18]。

- 少ないデータ数を持つクラスをオーバサンプリングしてもう一方のクラスのデータ数に合わせる。
- 多くデータ数を持つクラスをサブサンプリングしてもう一方のクラスのデータ数に合わせる。
- 一方のクラスを無視して、もう一方のクラスをカバーするような規則を獲得する。

我々が集めたデータセットは、ある程度の大きさのデータ量を持つので、2番目のデータをサブサンプリングする手法を用いた。なお、テストデータによる評価値算出の際は、サブサンプリングは行っていない。

また一般に、機械学習の性能は用いる素性数（素性に利用する語彙のサイズ）に依存する。そこで、素性数を100個から2,000個、そして素性数の最大値まで変化させて

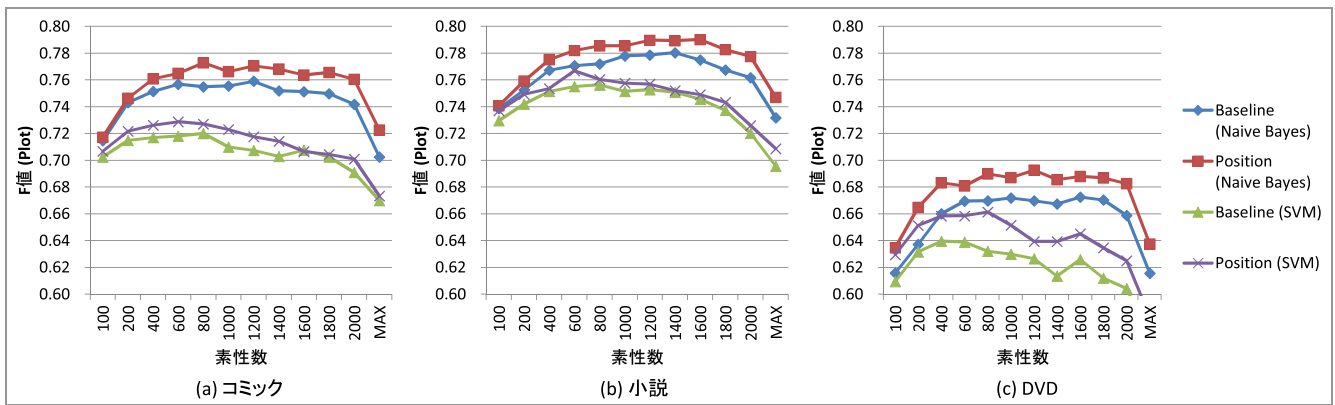


図 6 素性数を変化させたときの F 値の変化

Fig. 6 The change of F-value when changing the number of attributes.

評価を行った。素性数は、100-1,000 個までは 100 個おきに、1,000-2,000 個までは 200 個おきにとりうる素性数を変化させた。この際、学習に用いる単語は、Glover らの手法 [19] に従い、あらすじのラベルとの相互情報量の高い順に選択した。また、データセット中、出現回数が 1 回しかないものは、その相互情報量によらず、素性としては採用しないことにした。先行研究 [1] から、素性数が 100-2,000 個のときに Naive Bayes と SVM のどちらもピークを迎えることが分かっているため、この範囲を詳細に分析することにした。

本稿で使用した 5 つの学習アルゴリズムは、データマイニングツール Weka^{*1} を用いている。SVM は線形カーネルを用いており、パラメータ C (complexity constant) は 1 である。Logistic 回帰の入力値について説明する。スコアの学習アルゴリズムが Naive Bayes の場合は、Naive Bayes の確率値を用いている。0 から 1 の範囲の値をとる。スコアの学習アルゴリズムが SVM の場合は、SVM の出力に Logistic 回帰モデルを使用することで算出した確率値を用いている。0 から 1 の範囲の値をとる。

また、評価尺度として、あらすじである文の判定に対する F 値を用いる。F 値 (F-Value) は精度 (Precision) と再現率 (Recall) の調和平均により算出される指標であり、正確性と網羅性の総合的な評価の際に利用される尺度である。

7.2 素性数ごとの実験結果

各カテゴリにおいて素性数を変化させたときの実験結果を示す。本節では、ベースラインの手法を Baseline、位置情報を付与した場合を Position と表記する。

図 6 に、学習アルゴリズムに Naive Bayes を用いた場合の実験結果と、SVM を用いた場合の実験結果を示す。各手法において最も高い F 値の素性数を最適な素性数と呼ぶことにする。表 5、表 6 に各アルゴリズムの最適な素性

表 5 最適な素性数 (Naive Bayes)

Table 5 The appropriate number of attributes for Naive Bayes.

	コミック	小説	DVD
Baseline	1,200	1,400	1,600
Position	800	1,600	1,200

表 6 最適な素性数 (SVM)

Table 6 The appropriate number of attributes for SVM.

	コミック	小説	DVD
Baseline	800	800	400
Position	600	600	800

数を示す。図 6 から分かるように、Naive Bayes で比較的高い F 値となっている 1,000 から 2,000 の素性数や、SVM で比較的高い F 値となっている 400 から 1,000 の素性数の、どの素性数で固定して見たときにも、手法間の順序関係 (F 値の高さの順序) が入れ替わることは起きていない。また、各手法の最適な素性数で見たときの順序関係とも一致している。そのため、以降の実験では、位置情報を付与する場合としない場合それぞれで、各カテゴリの最適な素性数を用いて、文脈情報の利用と意見判定の利用の有効性を調査することにする。

DVD カテゴリに対する考察を行う。DVD カテゴリの F 値が他のカテゴリと比べて低いのは、複合的な要因である。1 つ目の要因は、3 人の評価者の判定の一致度が他のカテゴリより低いことである。判定の一致度が低いということは、判定が困難な文が多いということを示す。2 つ目の要因は、DVD カテゴリのあらすじクラスの含有率 (データセット中のすべての文のうち、あらすじクラスの付いた文の割合) が低く (表 1 参照)、ランダム判定での F 値が他のカテゴリより低いことである。ランダム判定とは、ここでは判定に何の妥当性もない手法として、ランダムにあらすじか否かを判定する手法のことを指す。この手法において、コミック、小説、DVD の F 値はそれぞれ 0.378, 0.399, 0.361 となる。3 つ目の要因は、DVD カテゴリのあ

*1 <http://www.cs.waikato.ac.nz/ml/weka/>

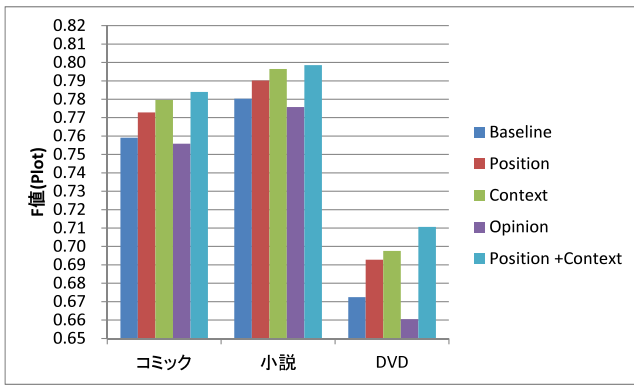


図 7 提案手法のあらすじ判定の F 値 (Naive Bayes)

Fig. 7 The F-value of the baseline and the proposed method in Naive Bayes.

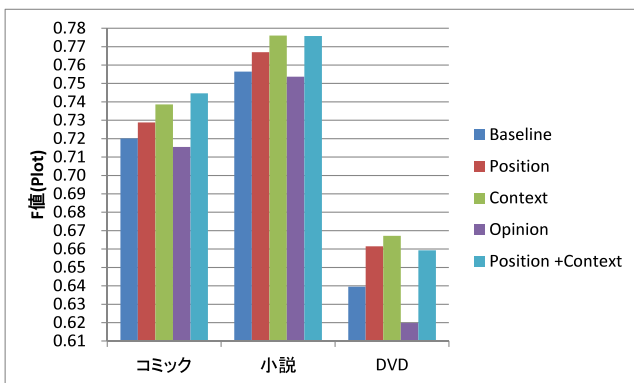


図 8 提案手法のあらすじ判定の F 値 (SVM)

Fig. 8 The F-value of the baseline and the proposed method in SVM.

らすじクラスの含有率が低いいため、サブサンプリングにより除去される学習データの量が多いためである。なお、1つ目の要因にあげた判定の一致度の低さは、この分類問題の難しさを示すとともに、正解データとしての信頼性の低さも示している。したがって、DVD カテゴリの精度の評価については、他のカテゴリよりも信頼性がやや劣ることも注意する必要がある。

7.3 提案手法の有効性

文書の構造情報として位置情報と文脈情報を用い、他クラスの情報として意見判定の結果を利用した場合に、あらすじ判定の性能がどれだけ向上するのかを調べる。本節では、位置情報を付与した場合を Position、文脈情報を利用した場合を Context、意見判定を利用した場合を Opinion と表記する。また、位置情報と文脈情報の両方を利用した場合を Position+Context と表記する。Context は図 5 の (c) で基本あらすじ分類器 (A) を用いる場合であり、Opinion は図 5 の (d) で基本あらすじ分類器 (A) を用いる場合であり、Position+Context は図 5 の (c) で基本あらすじ分類器 (B) を用いる場合である。

図 7, 図 8 に、学習アルゴリズムに Naive Bayes を用い

表 7 F 値・精度・再現率 (Naive Bayes)

Table 7 The F-value, precision and recall of the plot classification in Naive Bayes.

(a) コミック					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.76	0.77	0.78	0.76	0.78
精度	0.77	0.78	0.80	0.79	0.80
再現率	0.75	0.77	0.76	0.73	0.77
(b) 小説					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.78	0.79	0.80	0.78	0.80
精度	0.80	0.79	0.82	0.81	0.82
再現率	0.76	0.79	0.77	0.74	0.78
(c) DVD					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.67	0.69	0.70	0.66	0.71
精度	0.69	0.71	0.75	0.72	0.75
再現率	0.65	0.68	0.65	0.61	0.67

表 8 F 値・精度・再現率 (SVM)

Table 8 The F-value, precision and recall of the plot classification in SVM.

(a) コミック					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.72	0.73	0.74	0.72	0.74
精度	0.69	0.69	0.76	0.74	0.76
再現率	0.75	0.77	0.71	0.69	0.73
(b) 小説					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.76	0.77	0.78	0.75	0.78
精度	0.74	0.73	0.80	0.78	0.79
再現率	0.77	0.80	0.75	0.73	0.76
(c) DVD					
	Baseline	Position	Context	Opinion	Position+Context
F 値	0.64	0.66	0.67	0.62	0.66
精度	0.59	0.60	0.73	0.69	0.70
再現率	0.70	0.74	0.62	0.56	0.62

た場合の実験結果と、SVM を用いた場合の実験結果を示す。表 7, 表 8 に、Naive Bayes を用いた場合の F 値・精度・再現率と、SVM を用いた場合の F 値・精度・再現率を示す。どちらのアルゴリズムにおいても、Baseline と比較して、位置情報を付与した場合と、文脈情報を利用した場合に、全カテゴリで判定精度が向上した。t 検定 (一対の標本による平均の検定) を各交差検定によって算出されたそれぞれの F 値に対して行う。その結果、Baseline-Position 間、Baseline-Context 間、Baseline-Position+Context 間は、全カテゴリで判定精度が有意に向上した ($p < 0.05$)。また、

位置情報の付与と文脈情報の利用の両方を行った場合が、全カテゴリで最も判定精度が高くなった。一方で、意見判定を利用して最も判定精度が向上しなかった。以上から、レビュー文書の文書構造（位置情報、文脈情報）を利用することで、あらすじの判定精度が向上することが分かった。

7.4 意見判定の利用についての考察

この節では、意見判定を利用して最も判定精度が向上しない理由を明らかにする。そのために、意見判定によって算出されたスコアの代わりに、意見かどうかのラベル（意見の正解データ）を利用してあらすじ判定を行った。つまり、意見らしさのスコアを意見である場合は1、意見でない場合は0とし、意見判定の利用と同様の手法を用いて学習を行う。その結果、Naive BayesにおけるF値はコミックが0.779、小説が0.796、DVDが0.705となり、意見ラベルを利用したあらすじ判定はベースラインよりも判定精度が向上した。

この結果から、各文が意見かどうかはあらすじかどうかを判断する良い情報であったといえる。ここから、意見判定の利用があらすじ判定に有効でなかったのは、意見判定の精度があらすじ判定に寄与させるには十分でなかったからと考えられる。しかし、本研究で用いた意見判定において、Naive Bayesでの意見に対するF値は、コミックが0.822、小説が0.820、DVDが0.830となっている。これらは、既存研究の意見判定の結果と比べて低いわけではない。つまり、あらすじ判定に意見判定を利用するためには、意見判定が非常に高い精度でなければならないことが分かる。このことから、実運用で意見判定をあらすじ判定の精度向上に用いるのは難しいといえる。

7.5 被験者実験

被験者実験により、提案手法によって spoiler（ストーリーに関する記述であり、ユーザが知りたくなかった記述）がどの程度除去され、有用な文（商品購入の際に意思決定につながる文）がどの程度表示できているかを調査した。比較する手法は、提案手法、ベースライン、意見表示手法、ランダム手法の4つである。意見表示手法は、意見判定によって意見と判定された文を表示する手法である。ランダム手法は、遮蔽するか否かを一定確率で決定し、その確率を変化させて評価する手法である。被験者実験に利用するデータセットは小説カテゴリからランダムに選んだ10個のアイテムを用いる。被験者は7人であり、レビューの全文を見て、各文が spoiler かどうかと有用な文かどうかを判定した。どのような文を spoiler や有用な文と感じるかは人によって異なる。そのため、各被験者の判定それぞれを評価データと見なし、全被験者の平均値を評価値とする。

図9に各手法における spoiler の除去率とレビュー文の表示率の関係、図10に各手法における spoiler の除去率

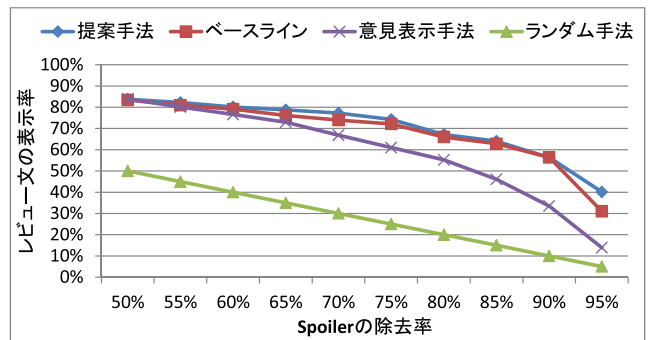


図9 各手法における spoiler の除去率とレビュー文の表示率の関係
 Fig. 9 The relationship among the spoiler hidden ratio and the ratio of the displayed review sentences.

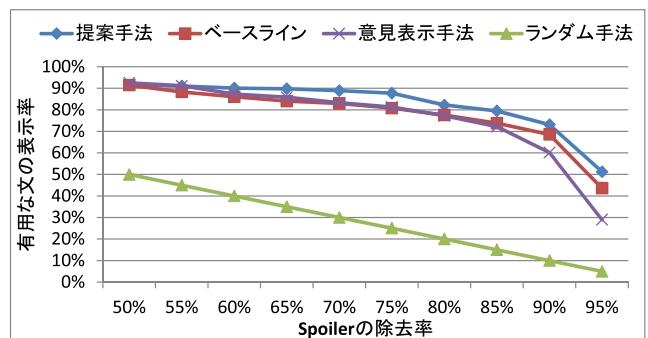


図10 各手法における spoiler の除去率と有用な文の関係
 Fig. 10 The relationship among the spoiler hidden ratio and the ratio of the displayed useful review sentences.

と有用な文の関係を示す。 spoiler の除去率は各被験者の spoiler と判定されたすべての文のうち、各手法で除去できた文の割合である。レビュー文の表示率はレビューのすべての文のうち、各手法で除去されなかった文の割合である。有用な文の表示率は各被験者の有用と判定されたすべての文のうち、各手法で除去されなかった文の割合である。 spoiler の除去率に対するそれぞれの値は、分類器の出力したスコアに対して文を除去するしきい値を設けて、そのしきい値を変化させることで算出した。それぞれの spoiler の除去率においてレビュー文の表示率が高いほど、 spoiler を除去してもなお元のレビューを維持できているため、より優れているといえる。また、それぞれの spoiler の除去率において有用な文の表示率が高いほど、 spoiler を除去してもなお有用な文を維持できているため、より優れているといえる。

図9から、あらすじ判定の手法は意見表示手法よりレビュー文の表示率において優れていることが分かる。また、図10から、提案手法は文書構造を考慮しない手法より有用な文の表示率において優れていることが分かる。よって、提案手法はユーザが知りたくなかった記述の多くを除去できており、かつ、商品購入の際に意思決定につながる文の多くを表示できているといえる。

表 9 文の除去箇所の個数 (Naive Bayes)

Table 9 The number of the places where sentences are deleted.

手法	コミック	小説	DVD
Baseline	682	627	680
Context	501	446	429

7.6 文脈情報の利用についての考察

文脈情報を利用した場合について詳しい調査を行う。

先行研究 [2] では被験者実験を用いて、文を除去してレビューを提示した場合に、人はどの程度レビューに違和感をいだくかを調査されている。ここで、その調査方法と結果の概要を示す。比較対象は、本研究における Baseline のあらすじ判定と、ランダム判定である。ランダム判定は、あらすじ判定で消去された文の数をランダムにあらすじと判定する。被験者は、あらすじと判定された文を消去して読んだレビューで、レビューの文脈に違和感をいだいた箇所を判定する。

あらすじ判定とランダム判定では、それぞれ同じ数の文を除去している。文の除去箇所の個数（文が除去された箇所の個数とし、連続して文が除去されたところも 1カ所と見なす）はそれぞれ、あらすじ判定は 39、ランダム判定は 77 である。被験者実験の結果、レビューの文脈に違和感をいだいた割合（レビューの文脈に違和感をいだいた回数を、表示した文の数で割った値）は、あらすじ判定 9.1%、ランダム判定 15.3% となった。ここから、レビューから同じ文の数を除去した場合でも、文の除去箇所の個数が多いほど、レビューに違和感をいだく割合が多くなるのが分かる。よって、レビューから同じ文の数を除去した場合には、その除去した文が連続している方が好ましい。

そこで、本研究では、文脈情報を利用した場合と利用しなかった場合のそれぞれで、文の除去箇所の個数を調査した。表 9 に、Naive Bayes において文脈情報を利用した場合の文の除去箇所の個数をベースラインと比較した結果を示す。文の除去箇所の個数は、あらすじを除去してレビューを表示した際に、文が除去された箇所（連続して文が除去されたところは全部で 1カ所とする）の個数とする。その結果、文脈情報を利用した場合、どのカテゴリにおいても文の除去箇所の個数を抑えられることが分かった。

調査からあらすじは連続しやすいという特性を持つことが分かっており、文脈情報を利用する手法はその特性を考慮して判定精度を向上させている。実際のレビューを調査すると、あらすじが連続しているなかで 1文だけ誤判定であらすじでない判定されていた文が正しく判定される、もしくはあらすじでない文が連続しているなかで 1文だけ誤判定であらすじであると判定されていた文が正しく判定されるといった事例が見られた。これらの判定は、あらすじを除去してレビューを見たときの読みやすさの向上に寄与すると思われる。

8. おわりに

本研究では、ストーリーを持つアイテムに対するレビューにおいて、レビュー文書中の文を、文中の単語のあらすじらしさだけでなく、その文が属する文書の文脈一貫性も用いて、あらすじか否かを判定する手法を提案した。文脈一貫性として文書の構造情報とあらすじ以外のクラス情報を利用する。構造情報としては、レビュー文書中の文の数（文数情報）、レビュー文書中での位置（位置情報）、前後の文のあらすじらしさ（文脈情報）に注目した。また、あらすじ以外のクラス情報としては、意見判定によって得られた意見か否かに関する情報（意見クラス）に注目した。

評価実験の結果、位置情報と文脈情報により、判定精度が向上することが確認できた。意見クラスも用いたが、これはあらすじ判定に貢献しなかった。被験者実験により、提案手法はユーザが知りたくなかった記述の多くを除去できており、かつ、商品購入の際に意思決定につながる文の多くを表示できていることが分かった。また、文脈情報を利用することで、あらすじである文を除去してレビュー文書を表示したときに、レビューの文脈に違和感をいだく割合が少なくなる可能性があることを示した。

今後は、文脈情報を利用してあらすじを除去した際に、レビュー文書の読みやすさが向上するかどうかを被験者実験により調査する予定である。

参考文献

- [1] 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).
- [2] 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, インタラクシオン 2013 論文集, pp.1-8 (2013).
- [3] Pang, B. and Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proc. ACL '04*, pp.271-278 (2004).
- [4] Yu, H. and Hatzivassiloglou, V.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, *Proc. EMNLP '03*, pp.129-136 (2003).
- [5] Riloff, E. and Wiebe, J.: Learning Extraction Patterns for Subjective Expressions, *Proc. EMNLP '03*, pp.105-112 (2003).
- [6] Dave, K., Lawrence, S. and Pennock, D.M.: Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, *Proc. WWW '03*, pp.519-528 (2003).
- [7] Wilson, T., Wiebe, J. and Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, *Proc. HLT '05*, pp.347-354 (2005).
- [8] Taboada, M. and Grieve, J.: Analyzing Appraisal Automatically, *Proc. AAIL '04*, pp.158-161 (2004).
- [9] 江崎晃司, 松井藤五郎, 大和田勇人: Weblog 上の評判情報における形容詞の出現位置を考慮した賛否分類, 第 67 回情報処理学会全国大会, No.2, pp.393-394 (2005).

- [10] Otsubo, M., Hijikata, Y. and Nishida, S.: Web Page Classification using Anchor-related Text Extracted by a DOM-based Method, *Trans. JSAI*, Vol.25, No.1, pp.37-49 (2010).
- [11] 那須川哲哉, 金山 博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会第 162 回自然言語処理研究会, pp.109-116 (2004).
- [12] 梅澤佑介, 乾 孝司, 山本幹雄: 評価表現と文脈一貫性を利用した教師データ自動獲得によるクレーム文検知, 電子情報通信学会技術研究報告, Vol.111, pp.47-52 (2012).
- [13] John, G.H. and Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers, *Proc. UAI '95*, pp.338-345 (1995).
- [14] Vapnik, V.N.: *Statistical Learning Theory*, Wiley-Interscience (1998).
- [15] Siegel, S. and Castellan, N.J.Jr.: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill (1988).
- [16] Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data, *International Biometric Society*, Vol.33, No.1, pp.159-174 (1977).
- [17] Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proc. IJCAI '95*, Vol.2, pp.1137-1145 (1995).
- [18] Japkowicz, N.: *Learning from Imbalanced Data Sets: A Comparison of Various Strategies*, AAAI Press, Vol.68, pp.0-5 (2000).
- [19] Glover, E.J. et al.: Using Web Structure for Classifying and Describing Web Pages, *Proc. WWW '02*, pp.562-569 (2002).



岩井 秀成

2012 年大阪大学基礎工学部システム科学科卒業。2014 年同大学大学院修士課程修了。同年トヨタ自動車株式会社入社。2013 年インタラクシオン 2013 ベストペーパー賞, WebDB Forum 2013 優秀論文賞各受賞。在学中

は, テキストマイニングの研究に従事。



土方 嘉徳 (正会員)

1996 年大阪大学基礎工学部システム工学科卒業。1998 年同大学大学院修士課程修了。同年日本アイ・ビー・エム(株)東京基礎研究所入社。2002 年より大阪大学大学院基礎工学研究科システム創成専攻助手。2009 年より

同准教授。2005 年インタラクシオン 2005 ベストペーパー賞, 2006 年 ACM IUI Best Paper Award, DEWS2006 優秀論文賞, 2011 年 WebDB フォーラム 2011 最優秀論文賞, 2012 年 WebDB フォーラム 2012 優秀論文賞, 2013 年インタラクシオン 2013 ベストペーパー賞, 情報処理学会山下記念研究賞, WebDB フォーラム 2013 優秀論文賞各受賞。情報推薦, Web インテリジェンス, テキストマイニングの研究に従事。人工知能学会, ヒューマンインタフェース学会, 日本データベース学会ほか会員。電子情報通信学会シニア会員。博士(工学)。



西田 正吾 (正会員)

1952 年 1 月 5 日生。1974 年東京大学工学部電子工学科卒業。1976 年同大学大学院修士課程修了。同年三菱電機(株)入社。同社中央研究所システム基礎研究部研究員, グループマネージャーを経て, 1995 年大阪大学基礎

工学部教授。その後, 大阪大学基礎工学研究科長・基礎工学部長, 理事・副学長を経て, 現在, 大阪大学基礎工研究科教授。システム技術, ヒューマンインタフェース技術, メディア技術の研究に従事。1984~1985 年 MIT メディアラボ客員研究員。ヒューマンインタフェース学会論文賞(2001 年, 2005 年), 電気学会業績賞(2004 年), 船井情報科学振興賞(2006 年)等受賞。IEEE Fellow。電子情報通信学会フェロー。電気学会フェロー。著書は, 「ヒューマン・コンピュータ交流技術」(オーム社, 共著), 「メディア工学」(朝倉書店), 「情報メディア工学」(オーム社, 共著)等。工学博士。

(担当編集委員 乾 孝司)