Context-based Plot Detection from Online Review Comments for Preventing Spoilers

Yoshinori HIJIKATA, Hidenari IWAI, Shogo NISHIDA
Graduate School of Engineering Science
Osaka University, Japan
Email: hijikata@sys.es.osaka-u.ac.jp

***Abstract*—The plot (content or storyline of a story) may disappoint users who read review comments associated with items containing a story such as comics, novels, and movies. This paper proposes a new method for identifying sentences that include descriptions of the plot of the story. Conventional methods only use information based on the words contained in a target sentence; however, our new method uses contextual information in addition to word information. We identify contextual information by using the sentence location and the plot probability of surrounding sentences. An experiment showed that this method improved the accuracy with which plot-related information can be identified.**

***Keywords*-opinion mining; spoiler; plot detection; context;**

I. INTRODUCTION

People can read comments other users have entered for specific items on online shopping sites or price-comparison sites. These comments help users decide which item to buy. However, review comments entered in response to items relating to a story, such as novels, comics, and movies, sometimes include part of the content of the story, also known as the plot. Some of these descriptions refer to the ending of the story or contain details of the storyline. For example, some reviewers reveal the name of the person who committed the crime or share details of the conspiracy in reviews of a mystery novel. However, when people watch movies or read novels, they usually enjoy to imagine what will happen in the next scene of the story [1], [2]. Therefore, these kinds of revelations might disappoint people because they remove some of the fun people derive from the content [3]. We can say that information about the plot in review comments might be spoilers for users [4].

Our research group has been developing a method for detecting information relating to plots in online review comments [5]. Our method judges whether a specific sentence includes story content. We expect to use the method to develop a system that browses review comments to hide those parts including information about the plot based on a sentence (hereinafter, "spoiler-hiding system"). The system therefore enables users to read review comments without parts including story content (See Figure 1 which depicts our spoiler-hiding system under development). The method employs techniques consisting of a bag-of-words model and machine learning algorithms. Furthermore, it has achieved high accuracy of detection by generalizing people's names and words peculiar to a specific item. However, it only uses information contained in the words of a target sentence (a sentence to be judged). The detection accuracy might be improved by considering the context of the target sentence in a review document (review comments added by one user in relation to an item).

Figure 1. Plot sentences are hidden by our spoiler-hiding system under development. The system is applied to review comments in amazon.com.

In this study, we focus on the context of the target sentence. We consider the context to be the role of the target sentence in a review document such as being the introduction, main body, or conclusive remark in the review comment. We also consider context to be the meaning or standpoint the target sentence maintains in the sequence of sentences. Concretely, we use two types of context: (i) the location of a target sentence in the review document and (ii) the probability of its surrounding sentences containing plot-related information (also referred to as its "neighborhood.") First, we investigate the relationship between context and plot to find important features to identify plot sentences. Then, we propose our new plot identification method incorporating the features we found. In this paper, when a sentence includes content from the story, we consider it to be part of the plot regardless of its importance (the level of detail or the impact of the content). In our work we only target reviews in the English language.

The remainder of this paper is organized as follows. First, we introduce some related work. Then, we introduce the baseline method we proposed in a previous study [5]. We also explain the possibility that context information might contribute to plot identification. After that, we introduce the dataset and investigate the relationship between the context and the plot. Then we propose our new plot identification method and conduct an experiment to assess its perfor-

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.19

mance. We give some discussions and show the limitation of our method based on the experimental results. Finally, we present some conclusions and suggestions for future work.

II. Related work

A. Spoiler filtering

Recent years have seen studies of spoiler detection or plot detection from review comments or articles in social media.

Golbeck tried to identify and block every tweet on a given topic [6]. She especially tried to block tweets on TV programs and sporting events. She detected spoiler tweets on TV programs by extracting actor and character names from the IMDB (Internet Movie Database). Spoiler tweets on sporting events were detected by the names of players, teams, and stadiums. Nakamura et al. developed a spoiler filtering system for sports news [4]. In this system, users are required to input several words relating to their favorite sports team or sporting event such as "Yankees" and "Olympic." Then, the system displays the news on the Web by hiding the results of the relevant sporting event. They also compared several interfaces such as deletion and blacking out to hide the results of sporting events [7]. These two systems used a rule-based method using given keywords. However, the method cannot be applied to review comments for items with story because they include various kinds of words related to a spoiler.

Guo and Ramakrishnan considered that spoilers of items, including stories, are related to plot descriptions (e.g. item descriptions in a shopping site) [8]. They deleted spoilers by calculating the similarity between the synopsis obtained from the item description page on the IMDB website and the user comment. The documents are presented in topics obtained by LDA. Because this method relies on general topics, it is difficult to apply this method to sentence-level spoiler (or plot) detection.

B. Using context information

In the area of sentiment analysis [9], it is well known that the use of location information of sentences improves the performance of opinion classification. Taboada et al. showed that descriptions containing the writer's opinion are not distributed equally throughout a review document and are found in a smaller part in the document [10]. They also improved the accuracy of opinion classification by changing the feature values of words according to the location of the word in the document. Otsubo et al. improved the classification accuracy of Web pages using the HTML document structure. They assigned a high weight to words that exist near the target anchor [11].

It is also known that the use of surrounding sentences in addition to the target sentence can improve the performance of opinion classification. This idea is derived from context coherence in which text spans occurring near each other tend to share the same subjectivity status [12]. Pang and Lee performed subjectivity detection on individual sentences and examined the relationships of subjectivity between two sentences [13]. Kanayama and Nasukawa showed that the polarity of a target sentence tends to be the same as that of surrounding sentences [14], [15]. They used this knowledge to identify polar clauses. Zhou et al. tried to obtain intra-sentence discourse relations by considering whether two text segments have the same or opposite polarity [16]. Inui et al. tried to detect sentences considered to be a claim made by the user [17]. They improved the accuracy of detection by assigning a high weight to sentences adjacent to a sentence identified as a claim. Recently, Yang and Cardie expanded this idea from a local constraint (adjacent sentences) to long-distance discourse relations [18].

The above work applied the sentence location and context information to opinion classification, web page classification, and claim classification, whereas our study applied this information to plot detection.

III. Research Approach

We introduce the baseline method we proposed in the previous study [5], the possibility that the context might contribute to plot identification, and the dataset we created for the evaluation.

A. Baseline method

The baseline method [5] applies machine-learning algorithms to plot identification. It judges each sentence of a new user review using a learned classifier as plot description or not. Sentence p is represented in a bag-of-words model as

$$p = \langle w_1, w_2, ..., w_M \rangle. \quad (1)$$

Therein, w_m represents a word. The number of occurrences $x_{n,m}$ of each word w_m in sentence p_n is also recorded. When creating the bag-of-words model, we did not remove stop words because some of these words may be related to the plot (Actually, we found personal pronouns that are related to the plot in [5]). The Porter stemming algorithm is implemented to remove morphological and inflectional endings from words. Previously, we used Naive Bayes, SVM, logistic regression, decision tree, and k-nearest neighbor as machine learning algorithms.

B. Usage of context information

We expect the baseline method to be improved by using context information in addition to word information. We explain this idea using the following example of a review document. Let us consider the last sentence "Some of the other ..." as a target sentence.

The previous study [5] investigated types of words highly relevant to 'plot' or 'non-plot'. The results showed that human names and their pronouns ("he" and "she") often occur in a plot sentence. This is because reviewers usually write "Who did it?" or "Who thought so?" when they refer

⟨An example of a review document⟩
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.

to the story. The target sentence in the above example is likely to be judged as a plot because it includes pronoun "she". However, it is not actually a plot sentence and the judgment becomes incorrect.

However, the probability to be judged as a plot is likely to decrease when the location of the sentence in the document and the probability of its surrounding sentences containing the plot are considered. Sentences located near the top or bottom of the review document are not likely to relate to the plot. This is because reviewers usually give priority to writing their impressions of the item rather than writing the content of the story. They usually conclude their review comments with a summary of the review. The target sentence appears last, therefore the probability to be judged as plot becomes lower.

It is also likely that plot sentences (or non-plot sentences) occur in sequence in the document. A sentence is highly likely to be a plot when the surrounding sentences are related to the plot. Readers find sentences with frequent changes of context hard to read. Therefore, people write sentences with contextual consistency. In the above example, the sentence before the target sentence is not a plot (the sentence is likely to be judged as non-plot by a machine learning algorithm). Thus, the target sentence is also likely not a plot.

In this study, we clarify the correctness of the above supposition in Section IV.

C. Dataset

We used the same dataset of review comments we previously created in [5]. This dataset contains review comments from amazon.com in the comic, novel, and movie categories, in which each sentence is labeled as being related to either the plot or the non-plot. A single user's review comment to an item (review document) is a unit of collection. All the sentences in one review document are recorded in the database. One dataset of each category contains 500 review documents randomly selected from amazon.com. Three human evaluators assigned labels to each sentence according to whether the sentence includes the content of the story (considered as plot). We treated that sentence as belonging to 'plot' class if two or more of three people regarded a certain sentence as plot.

Table I shows the number of words, words occurring more than once, sentences assigned to the 'plot' class, and sentences assigned to the 'non-plot' class in the dataset.

Table I
STATISTICAL DATA OF THE DATASET IN EACH CATEGORY

	comic	novel	movie
♯ words	6414	6334	6966
♯ words ($\geq$2)	3603	3539	3761
♯ plot sentences	1523	1602	1357
♯ non-plot sentences	3484	3225	3445

Table II
EXAMPLE OF A REVIEW DOCUMENT AND SENTENCE LOCATION.

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

The number of words refers to the number of kinds of words that correspond to elements of the bag-of-words. We calculated the kappa coefficient [19] to ascertain the degree of accordance of labels among the three evaluators. The results of plot labels are 0.612 for the comic category, 0.544 for the novel category, and 0.466 for the movie category. Generally, the value of the kappa coefficient is regarded as low when the degree of accordance is in the range 0–0.4, as medium in the range 0.4–0.6, as high in the range 0.6–0.8, and as extremely high in the range 0.8–1.0 [20]. The labels acquired in our data set preparation have medium or high accordance.

IV. INVESTIGATION

A. Sentence location

We investigate the relationship between the location of each sentence in the review document and its class in terms of plot. Table II contains an example of a review document and the location numbers of their sentences. Sentence location is defined as the n-th sentence from either the head or end of the document (a smaller value of n is selected). We examined the ratio of plot sentences to the sentences in each location. We separated the locations nearer to the head of the document ("first half")) from those nearer to the end of the document ("last half"). In detail, the second sentence "She has created a wonderful world ..." in Table II is to be included in the first half. The fourth sentence "This is one book I know I'm going to ..." in Table II is to be included in the last half".

Figure 2 shows the ratio of the number of plot sentences to all the sentences in each location. When the document contains an odd number of sentences, the middle sentence is excluded from this examination to ensure that the difference between the first half and last half is determined correctly. The number of sentences in each location is depicted on the bar graph. The ratio of plot sentences is depicted in the line graph. Because the line graphs increase steadily, sentences located far from the head (end) of the document

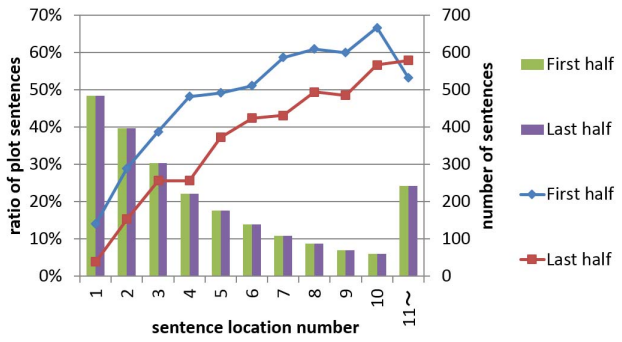

Figure 2. Ratio of plot sentences (line graph) and the number of sentences (bar graph) for each sentence location. Sentence location is counted from the head of the document (blue line and green bar) or the end of the document (red line and purple bar).

tend to include the content of the story. Because the line graph of the first half (blue line) lies above that of the last half (red line) in the graph, the first half includes more plot descriptions than the last half. From this result, we decided to use the location information (hereinafter "sentence location information") to identify the plot. We also distinguish the first half from the last half. When the document contains an odd number of sentences, the middle sentence is included in the first half in our plot identification method.

B. Neighborhood plot probability

We examine whether surrounding sentences affect the class of the target sentence. The system can infer the classes of not only the target sentence but also its surrounding sentences. We can use the inferred classes of the surrounding sentences for inferring the class of the target text. We refer to this information as "neighborhood plot probability." Actually, we investigate the relationship between the target sentence and its surrounding sentences to determine whether the class is plot or non-plot. This investigation checks one or two sentences located before and after the target sentence. We categorize the set of surrounding sentences into a specific sequential pattern of classes (e.g., 'non-plot' - 'plot' - 'target sentence' - 'plot' - 'plot'). We then calculate the ratio of plot sentences to the target sentences in each class pattern.

Table III lists the ratio of plot sentences and the frequency in the dataset in each class pattern of surrounding sentences. Note that 'X' presents the target sentence. A class pattern is denoted using 'T' and 'F'. If a sentence is assigned to the 'plot' class, it is denoted as 'T'. If it is assigned to the 'non-plot' class, it is denoted as 'F'. For example, if the second-former sentence is assigned to the 'non-plot' class, the first-former sentence is assigned to the 'plot' class, the first-latter sentence is assigned to the 'plot' class, and the second-latter sentence is assigned to the 'plot' class, the pattern is denoted as "FTXTT". The left part of the table shows a

Table III
RATIO OF PLOT SENTENCES IN EACH SEQUENTIAL PATTERN OF CLASSES.

Surrounding two sentences			Surrounding four sentences		
Pattern	Ratio	Freq	Pattern	Ratio	Freq
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

Freq: The number of occurrences of the pattern

case in which only one former and one latter sentence are considered ("Surrounding two sentences" in the table). The right part of the table shows a case in which two former and two latter sentences are considered ("Surrounding four sentences" in the table).

This result shows that when both the former one sentence and the latter one sentence are assigned to the "plot" class, 90% of the target sentences are plot. On the other hand, when they are assigned to the "non-plot" class, only 9% of the target sentences are plot. This result indicates that the adjoining sentences are strongly related to the type of target sentence (plot or non-plot). The ratio becomes more biased when considering one additional former and latter sentence. Based on this result, we use the surrounding four sentences (the two adjoining sentences before the target sentence and those after the target sentence) for plot identification.

V. PLOT IDENTIFICATION METHOD

This section explains the baseline method which was proposed in [5] and our proposed method which is an improved version of the baseline.

A. Baseline method

The basic principles of the baseline method [5] were explained in Subsection III-A. We explain its detail here.

The procedure of the baseline method is the followings. A plot classifier was trained in advance and a new sentence will be assessed to determine whether it is related to the plot. We name this plot classifier "BPC (basic plot classifier)," which is also used for our proposed plot identification method. This procedure trains a plot classifier using sentences that are presented in bag-of-words model. Each sentence is assigned class information relating to the plot. Previously, we tested several kinds of machine learning algorithms, and we found

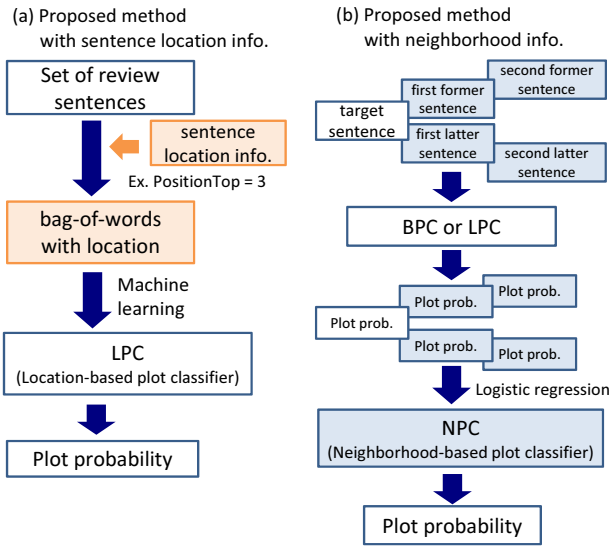

Figure 3. Flow diagram comparing the baseline method and our proposed method

Naive Bayes and SVM to outperform the other algorithms. This prompted us to use these two algorithms in this study.

The method should segment a document into sentences. The basic idea of segmentation is to search for the punctuation marks '.', '!' and '?' followed by the space character or line feed character. The end of a sentence can also be identified if any of the above three characters accompany a right bracket, or single or double quotation marks. We exclude cases with a special meaning such as "i.e.", "P.S." and "etc." when using '.' for segmentation.

The baseline method generalizes human names (character name, author name, and other human name) and peculiar words (words that have only occurred in review comments to a specific item) for improving the identification accuracy. Because human names and peculiar words either do not occur or rarely occur in several items when they are used in the original form, they do not perform well for plot identification. Thus, the baseline methods generalize these words by replacing them with tags such as ⟨character⟩ and ⟨peculiar⟩.

Details of the method for generalizing human names and peculiar words are provided in [5]. We briefly explain the method in this paper. Human names are generalized by identifying words which are registered in the database of names of people provided by the U.S. Census Bureau [1]. Among the found words, we excluded words that are also registered in a general English dictionary [2] (e.g., 'White' and 'Hill'). Words that also occur in an item description document [3] were identified as character names. Words defined as an author name in the above document were identified as author name. We generalized peculiar words by counting the number of items that received review comments including this word and we did this for each type of word. If there was only one item, we identified the word as peculiar word.

[1]http://www.census.gov, 1995 edition

[2]We used the online dictionary provided by ALC Press Inc.

B. Our proposed method

Our proposed method is an improved version of the baseline method and incorporates sentence location information and neighborhood plot probability. Figure 3-(a) shows the procedure of our proposed method when using sentence location information. This method learns a classifier, which classifies a sentence as being either plot or non-plot, using word information and sentence location information. We name the learned classifier "LPC (location-based plot classifier)," which will be also used for another version of our proposed method.

The detail of the first version of our method (using only sentence location information) is explained as follows. A sentence is presented as a word vector. The location of the sentence is added to another element of this vector. In detail, two new elements are added to the vector: an element presenting the location from the beginning of the document and that presenting the location from the end of the document. Only one element is assigned a value depending on whether the target sentence is nearer to the beginning or the end of the document. For example, if the sentence is near the top of the document and is the third one from the top, $PositionTop = 3$ is added as new element in the vector ($PositionEnd = 0$ here). We conducted binning here because the number of sentences with a high location value is small. Concretely, 11 is added to the above element when the location exceeds 11. A Naive Bayes or SVM algorithm is used to learn the classifier.

Figure 3-(b) shows the procedure followed by our proposed method when using neighborhood plot probability. This method uses a target sentence and the four sentences surrounding it. It calculates scores representing the probability to be a sentence relating to the plot for these five sentences using BPC or LPC. The target sentence is presented by the five scores obtained in the above manner. Finally, it learns another classifier (denoted as "NPC (neighborhood-based plot classifier)") by using the five scores and the label (plot or non-plot) of the target sentence. NPC is built using logistic regression. In the later experiment, data for building BPC (or LPC) and that for building NPC are separated in advance (the detail of data segmentation is provided later).

VI. EXPERIMENT

This section evaluates our plot identification method. Especially, we clarify whether sentence location information

[3]We used the item description pages on amazon.com

and neighborhood plot probability improve the accuracy of plot identification. We compare the following methods.

- *Baseline*: A plot identification method which uses only word information (uses only BPC). This is the same method proposed in [5].
- *Location*: Our proposed plot identification method which uses only sentence-location information (uses only LPC).
- *Neighborhood*: Our proposed plot identification method which uses only neighborhood plot probability (incorporating class information of surrounding sentences based on BPC).
- *Hybrid*: Our proposed plot identification method which uses both sentence-location information and neighborhood plot probability (incorporating class information of surrounding sentences based on LPC)

A. Learning condition

Ten-fold cross validation [21] is used for the evaluation. Because the proposed method builds classifiers in a hierarchical manner, we segmented data into the following three types: (i) a learning set for building BPC or LPC (using eight data segments among the ten segments. Hereinafter, referred to as "Learning set A"), (ii) a learning set for building NPC (using one data segment among the ten segments. Hereinafter, referred to as "Learning set B"), and (iii) a test set for the evaluation (using one data segment among the ten segments). The unit used in data segmentation is one item.

The detail of the cross validation is as follows. First, we build BPC or LPC using Learning set A and labels relating either to plot or non-plot. Then, we obtain the plot score for each sentence using the above BPC or LPC and Learning set B. We build NPC using the obtained plot scores and correct labels (plot or non-plot). Finally, we test the above classifiers using the test set. Note that nine data segments (Learning set A and Learning set B) are used to build PBC or LPC in *Baseline* and *Location* for keeping consistency of the overall learning set size.

Generally, the performance of a machine learning method is influenced by the number of attributes (the kinds of words in our study). We changed the number of attributes for use in machine learning from 100 to 2000 (and the maximum number of attributes) for the evaluation. We changed the number of attributes in increments of 100 from 100 to 1000, and in increments of 200 from 1000 to 2000. We selected words used as features based on the mutual information of a word and class (plot or non-plot), similar to the approach of Glover et al. [22]. We did not use words that only appeared once in the data set.

It should also be noted that imbalanced data might have a negative influence on the classification performance. As shown in Table I, the number of sentences in the 'plot' class is smaller than those in the 'non-plot' class in our data set. The following three measures are well-known against imbalanced data [23]: (i) conduct over-sampling of the data with a class of fewer data, (ii) conduct subsampling of the data with a class of larger data and (iii) ignore one of the two classes using a recognition-based instead of a discrimination-based inductive scheme. We used subsampling because our dataset is sufficiently large.

We used the datamining tool Weka for the experiment. The parameter settings of the Naive Bayes and SVM algorithms are as follows. We used a linear kernel for the SVM algorithm, where the value of parameter C (complexity constant) is 1. For the Naive Bayes algorithm, we used the probability score (which runs from 0 through 1) of the Naive Bayes as the plot score. In the case of SVM, we used the probability score (which runs from 0 through 1) which can be obtained by applying the logistic regression model to the outputs of SVM as the plot score.

We measure the accuracy of the classification. As accuracy metric, we use F-value on the plot class. The F-value reflects both the precision and recall of classification. Given the set of sentences P that are classified as belonging to the 'plot' class by the classifier and the set of sentences G that are assigned a 'plot' label in the ground truth data, the F-value is calculated as follows.

$$Precision = \frac{|P \cap G|}{|P|} \quad (2)$$

$$Recall = \frac{|P \cap G|}{|G|} \quad (3)$$

$$F\text{-}Value = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

B. Setting the number of attributes

Before evaluating our proposed method, we need to decide how many attributes (words) to use to learn the classifier (BPC for *Baseline* and *Neighborhood*), and LPC for *Location* and *Hybrid*). Then, we determine the best setting for the number of attributes by simply testing *Baseline* and *Location* here. The same numbers will be used for subsequent experiments for *Neighborhood* and *Hybrid*, which are explained later. Note that we used all the attributes of the location information in *Location* and *Hybrid*.

Figure 4 shows plots of the F-values of *Baseline* and *Location* as a function of the number of attributes. We decide to use the number of attributes that result in the maximum F-value. Tables IV and V list the number of attributes when the maximum F-value is obtained for Naive Bayes and SVM. We use the number of attributes shown in these tables for the later experiment.

Figure 4 indicates that the F-values for movie are less than those for comic and novel. This may be due to the following reasons. First, the inter-rater reliability for movie is less than those for comic and novel. This means that there is a larger number of sentences which are difficult to be categorized as either plot or non-plot. Second, movie has fewer plot sentences than comic and novel. Because

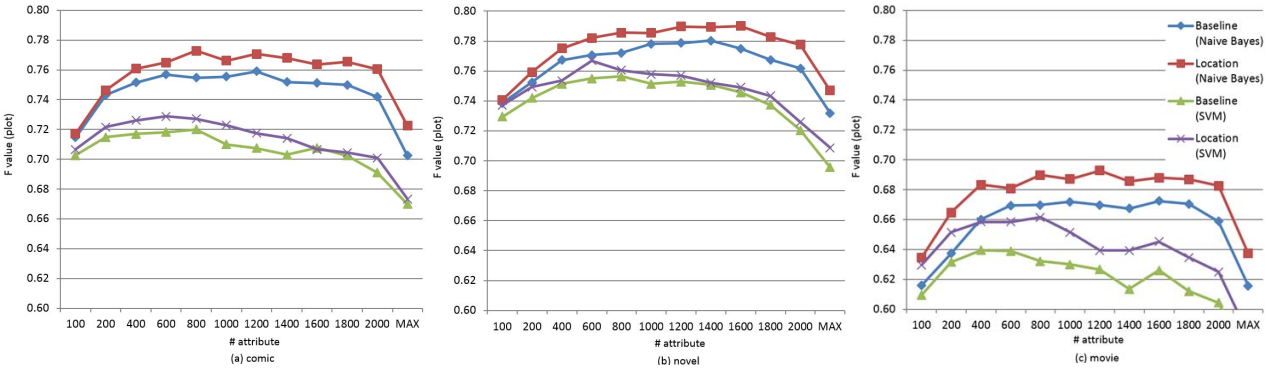

Figure 4. Changing the F-value when changing the number of attributes

Table IV
NUMBER OF ATTRIBUTES WHEN ACHIEVING THE MAXIMUM F-VALUE FOR NAIVE BAYES

	Comic	Novel	Movie
Baseline	1200	1400	1600
Location	800	1600	1200

Table V
NUMBER OF ATTRIBUTES WHEN ACHIEVING THE MAXIMUM F-VALUE FOR SVM

	Comic	Novel	Movie
Baseline	800	800	400
Location	600	600	800

we conducted subsampling, more sentences assigned to the 'non-plot' class are deleted when learning the classifier in movie. Therefore, the total number of sentences used for learning the classifier were decreased. We think that this decreased the F-value.

C. Validating our techniques

In this subsection, we show how much the performance improves when introducing the sentence location information and neighborhood plot probability for plot identification. We compare *Baseline*, *Location*, *Neighborhood* and *Hybrid*. Tables VI and VII list the F-values using Naive Bayes and SVM (also showing precision and recall). The F-values are improved when using sentence location information and neighborhood plot probability compared with *Baseline* in both cases using Naive Bayes and SVM. We conducted a student's t-test for the F-values obtained in each cross validation. There is significant differences between *Baseline* and *Location*, *Baseline* and *Neighborhood*, and *Baseline* and *Hybrid* ($p < 0.05$).

From this result, we can see that sentence location information and neighborhood plot probability contributes to the plot classification. Because we do not obtain a significant difference in the F-values between *Location* and *Neighborhood*, their contributions to the classification performance are almost same. However, the precision of *Neighborhood* is higher than that of *Location*, and the recall of *Neighborhood* is lower than that of *Location*. We can see that using the class information of the surrounding sentences leads to careful identification of plots. Furthermore, we can see that the F-values become the highest except movie domain when incorporating both sentence location information and neighborhood plot probability.

Table VI
F-VALUE, PRECISION AND RECALL (NAIVE BAYES)

(a) Comic

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.759	0.773	0.780	0.784
Precision	0.772	0.780	0.800	0.802
Recall	0.747	0.766	0.761	0.767

(b) Novel

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.780	0.790	0.796	0.799
Precision	0.799	0.795	0.822	0.817
Recall	0.762	0.785	0.773	0.780

(c) Movie

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.672	0.693	0.698	0.711
Precision	0.692	0.705	0.751	0.755
Recall	0.654	0.681	0.651	0.672

D. Detection example

Table VIII shows an example of plot detection when using *Baseline* and *Hybrid*. The table shows only the six sentence from the top of the review document. The second column shows the label about plot (ground truth data). The third and the forth column shows the class about plot classified by *Baseline* and *Hybrid* respectively. "P" shows the "plot" and "N" shows the "non-plot". Although, the fifth sentence is incorrectly identified as "non-plot" in *Baseline*, it is correctly identified as "plot" in *Hybrid*. This is because all of the surrounding sentences are identified as "plot" by the method. We can see that context information helped the method to

Table VII
F-VALUE, PRECISION AND RECALL (SVM)

(a) Comic

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.720	0.729	0.739	0.745
Precision	0.693	0.690	0.765	0.759
Recall	0.749	0.772	0.714	0.731

(b) Novel

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.756	0.767	0.776	0.776
Precision	0.740	0.734	0.802	0.792
Recall	0.773	0.803	0.752	0.760

(c) Movie

	Baseline	*Location*	*Neighborhood*	*Hybrid*
F-value	0.640	0.661	0.667	0.659
Precision	0.587	0.598	0.729	0.698
Recall	0.702	0.739	0.616	0.625

find a sentence related to the plot. However both *Baseline* and *Hybrid* cannot identified the second sentence as "plot". This sentence includes the word "author", which is strongly related to "non-plot" in general. It also exists near to the top of the review document. *Hybrid* cannot identify the sentence as "plot" due to the above reasons.

E. Discussion

The experimental results showed that both sentence location information and plot probability of the surrounding sentences contributes to the classification performance. We also found that the method carefully classifies a sentence to 'plot' class when using the surrounding sentences. We consider the reason lies in that some specific sequential patterns of classes like "TTXTT" and "FTXTT" are highly related to 'plot' class of the target sentence (See Table III). On the other hand, the ratio of plot sentences of each sentence location is about 60% at the maximum (See Figure 2).

The limitation of our method is that we simply use word information and do not conduct dependency parsing. Therefore, the method might judge a non-plot sentence as 'plot' incorrectly when the sentence includes a word which is highly-related to 'plot' (See the second sentence starting with "It was famous author" in Table VIII). In the proposed method, we did not use topics which can be obtained by LDA (Latent Dirichlet Allocation). The automatically-detected topics might include features regarding plot or non-plot. There is a possibility that applying leading-edge techniques of natural language processing solve this problem.

VII. CONCLUSION

This paper proposes a new method to identify sentences related to the plot in a document in which an item associated with a story is reviewed. Our method does not only use words as its source of information but also uses contextual information from the target sentence. In particular, we used the location of the target sentence in the review document and the probability of surrounding sentences in its immediate neighborhood containing information relating to the plot. Here, the sentence location refers to the number of sentences between the target sentence and the head or tail of the review document, whereas the plot probability refers to the probability of being related to the plot as inferred by the method.

An experiment showed that both location information and neighborhood plot probability contribute to the identification accuracy. Comparing the location information and neighborhood plot probability, utilizing neighborhood plot probability detects plots more carefully than utilizing the location information. Combining the location information and neighborhood plot probability achieves the highest performance. We hope that our method keeps users away from spoilers.

Future work aims to determine the kind of contents constituting major spoilers in each item that will prevent spoilers from being displayed to each user. We plan to utilize story content (e.g., the body text of a novel) to detect spoilers.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for challenging Exploratory Research 15K12150.

REFERENCES

[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] Nakamura, S. and Komatsu, T.: Temporal Filtering System to Reduce the Risk of Spoiling a User's Enjoyment, *Proc. of IUI '07*, pp. 345–348 (2007).

[5] Iwai, H., Hijikata, Y., Ikeda, K. and Nishida, S.: Sentence-based Plot Classification for Online Review Comments, Proc. of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence (IEEE/WIC/ACM WI'14), pp. 245–253 (2014).

[6] Golbeck, J.: The Twitter Mute Button: a Web Filtering Challenge, *Proc. of CHI '12*, pp. 2755–2758 (2012).

[7] Nakamura S. and Komatsu, T.: Study of Information Clouding Methods to Prevent Spoilers of Sports Match, *Proc. of AVI '12*, pp. 661–664 (2012).

[8] Guo, S. and Ramakrishnan, N.: Finding the Storyteller: Automatic Spoiler Tagging Using Linguistic Cues, *Proc. of COLING '10*, pp. 412–420 (2010).

Table VIII
EXAMPLE OF PLOT DETECTION

Sentence	Label	*Baseline*	*Hybrid*
The book Dying to Meet You: 43 Old Cemetery Road is an interesting thriller filled with surprises and humor.	N	N	N
It was famous author Ignatius B. Grumply's first day in the town of Ghastly.	P	N	N
He was looking for a new house to rent for the summer, so he started talking to a woman named Anita Sale from Proper Properties to help him find a home.	P	P	P
He was looking for a quiet house with no kids around, to write his newest Ghost Tamer book, but he was surprised when he overlooked the contract and ended up with a home that included a boy named Seymour and his cat.	P	P	P
The two lived on the third floor with their ghost friend, Olive C. Spence.	P	N	P
When Seymour's parents, Les and Diane Hope, can't find anyone to buy their haunted mansion, they plan to demolish it and sell it as an empty lot, but Seymour doesn't want this to happen so he struggles to make the money.	P	P	P

P: plot, N: non-plot

[9] Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, Journal Foundations and Trends in Information Retrieval, Vol.2, No. 1-2, pp. 1-135, (2008)

[10] Taboada, M. and Grieve, J.: Analyzing Appraisal Automatically, Proc. of AAAI'04, pp.158–161, 2004.

[11] Otsubo, M., Hijikata, Y. and Nishida, S.: Web Page Classification using Anchor-related Text Extracted by a DOM-based Method, Transactions of the Japanese Society for Artificial Intelligence, Vol.25, No.1, pp.37–49 (2010).

[12] Wiebe, J.M.: Tracking point of view in narrative, Computational Linguistics, Vol. 20, No. 2, pp. 233–287, (1994).

[13] Pang, B. and Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proc. of ACL'04*, pp. 271-278 (2004).

[14] Kanayama, H. and Nasukawa, T.: Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis, *Proc. of EMNLP'06*, pp. 355-363, (2006).

[15] Kanayama, H. and Nasukawa, T.: Unsupervised lexicon induction for clause-level detection of evaluations, Natural Language Engineering, pp. 83-107 (2012).

[16] Zhou, L., et al.: Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities, *Proc. of EMNLP'11*, pp. 162-171 (2011).

[17] Inui, T., Umezawa, Y., and Yamamoto, M.: Complaint Sentence Detection via Automatic Training Data Generation using Sentiment Lexicons and Context Coherence, Journal of natural language processing, Vol.20, No.5, pp. 683–705 (2013).

[18] Yang, B. and Cardie, C.: Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization, *Proc. of ACL'14*, pp. 325-335, (2014).

[19] Siegel, S. and Castellan, N.J.Jr.: Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill (1988).

[20] 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).

[21] Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proc. of IJCAI'95*, Vol.2, pp.1137–1145 (1995).

[22] Glover, E.J. et al.: Using Web Structure for Classifying and Describing Web Pages, *Proc. of WWW'02*, pp.562–569 (2002).

[23] Japkowicz, N.: Learning from Imbalanced Data Sets: A Comparison of Various Strategies, AAAI Press, Vol.68, pp.0–5 (2000).