Check for updates

Journal of KIIT. Vol. 20, No. 3, pp. 35-46, Mar. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 35
http://dx.doi.org/10.14801/jkiit.2022.20.3.35

Rethinking ROUGE Scores for Video Game Review Summarization

Yeong-Hun Lee*, Yuchul Jung**

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant (2020R1A4A1017775).

Abstract

Recall-oriented understudy for gisting evaluation (ROUGE) is a prevalent evaluation measure in the field of natural language processing, especially for text summarization. However, ROUGE's reliability has been continuously debated because a high ROUGE score does not guarantee a high-quality summary and vice versa. As an empirical study in the video game review summary, we address that existing state-of-the-art summarization techniques fail in generating high-quality game review summaries, and ROUGE scores for those summaries are quite problematic. To this end, game review data are newly collected, and BERT based automatic review summarizations are performed on the dataset to reconsider ROUGE use in video game review summarizations. Especially, we provide an in-depth discussion between the ROUGE score and the scores of human annotators in terms of game review factors.

요 약

Recall 기반의 요약 평가 지표(ROUGE)의 경우, 자연어처리 연구 분야 중 문장 요약 분야에서 널리 쓰여왔다. 그러나 ROUGE의 신뢰성에 대해서는 의문을 표해왔다. 그 이유로는 높은 ROUGE 점수를 받았다고 하더라도, 요약 결과의 품질이 좋지 못한 경우가 있었기에 ROUGE 점수가 요약의 품질을 보장해주지 못하였다는 점 때문이다. 본 연구에서는 그중 비디오 게임 리뷰 요약 분야에 집중하였다. 우리는 기존의 SOTA(state-of-art) 요약 기법들이 고품질의 게임 리뷰 요약 생성에 어려움을 겪고, 그 요약에 대한 ROUGE 점수가 문제가 있음을 확인하였다. 이를 위해서, 게임 리뷰 데이터를 새로 수집하고, 비디오 게임 리뷰 요약에 ROUGE 점수의 실효성을 확인하기위해 새로 구축한 데이터를 활용하여 BERT 기반의 자동 리뷰 요약 생성을 시도하였다. 그리고 게임 리뷰 요소 측면에서 ROUGE 점수와 사람이 직접 평가한 점수를 기반으로 분석을 진행하였다.

Keywords
summarization, ROUGE, game review, BERT

Ⅰ. Introduction

The video games market has risen steeply in 2020 compared to the previous years after the advent of COVID-19 pandemic [1]. This rapid growth rate has led to many games being released on the market. When shopping for a game that suits their preferences, users commonly read game review articles.

Several review articles are available from a wide variety of sources for a given game. However, the

* Graduate Student, Department of Computer Engineering, Kumoh National Institute of Technology
- ORCID: https://orcid.org/0000-0002-0750-2922
** Assistant Professor, Department of Artificial Intelligence Engineering, Kumoh National Institute of Technology
- ORCID: https://orcid.org/0000-0002-8871-1979

· Received: Nov. 01, 2021, Revised: Dec. 24, 2022, Accepted: Dec. 27, 2021

· Corresponding Author: Yuchul Jung
Department of Artificial Intelligence Engineering, Kumoh National Institute of Technology, 61 Daehak-ro, Geoui-dong, Gumi-si, Gyeongsangbuk-do
Tel.: +82-54-478-7536, Email: jyc@kumoh.ac.kr

form and length of these reviews can vary, and therefore, referring to multiple reviews can incur a significant time investment.

Several approaches exist for more accurate automatic text summarization, including extractive and abstractive approaches, in addition to the use of both in combination. Recall-oriented understudy for gisting evaluation (ROUGE) [2] is the most popular evaluation metric in text summarization although several alternatives are introduced. Meanwhile, the reliability of evaluations using ROUGE scores is debatable as it is based on recall. Several studies indicated the limitations of ROUGE and analyzed its shortcomings [3][4]. The most concerning problem with ROUGE's is its lack of ability to handle synonyms and identify terms related to specific topics or content. Moreover, the symptom becomes worse in video game review summarization because game users are very keen on game-related issues, such as sound, graphics, system, etc.

To investigate the feasibility of ROUGE metric for game review summarization, the contributions of our study are three folds:

1) A new game review and summary dataset was created using automatic crawling with a manual refinement process.
2) We performed game review summarization with three different BERT-large-cased models: BERT (transformer) [5], MatchSum [6], and BertSumExtAbs [7]), and analyzed them to identify the limits of the ROUGE score.
3) Comparative analysis between game specific quality factors and ROUGE scores was performed using sample reviews selected for five different games.

II. Related Work

Game Review Summarization: To perform a summarization task in the NLP domain, the dataset in the game field is very scarce. Most text summarization research studies utilize literary data or news data available in the public domain. For literature or news data, the sources used to build data sets vary but are very limited for games. The GameWikiSum dataset [8] that utilizes Wikipedia and game-related sites to build datasets was released recently. Since game reviews also have similar properties to other product reviews [9], clustering techniques by aspect and sentiment [10] and Double Propagation (DP) techniques based on aspect [11] have been studied.

Automatic Summarization: Recently, with the advent of BERT models, utilize BERT in many NLP fields. Following this trend, some studies are optimized for automatic summarization, such as BertSum [7]. The corresponding work adds [CLS] tokens at the beginning of the sentence and presents a method that combines an extractive technique with an abstract. And using the suggested methods, the results showed high performance.

RoBERTa [12] models with improved BERT models are also utilized in the field of automatic summarization. RoBERTa-based methods are trained in masked language modeling (MLM). MLM is useful for NLU (classification, regression, etc.) related tasks; however, it is inefficient for generation tasks.

More recently, an extractive technique, MatchSum framework [6] showed excellent performance on CNN/DailyMail datasets. For the comparison with SOTA level algorithms, we employ BertSum and MatchSum models to the field of game review summarization.

III. Dataset Construction

To construct a game review/summary dataset, we performed thorough data collection, game selection, and review analysis. To this end, game-related review texts were collected automatically from the review aggregation website, Metacritic. Through the screening process, five games were selected out of a total of 5,033 games. Finally, the reviews were narrowed down based on those containing summary statements for the games considered in the study; these reviews were then included in our final game review/summary dataset and used in our experiments.

This website offers high reliability in terms of video game-related data, and its metascore evaluation index has a considerable influence on game evaluation. As of March 25, 2020, the site aggregated reviews for a total of 5,033 games. We collected all reviews—a total of 278,849 reviews—and each item of the collected content comprised the game title being reviewed, author ID, review content, user score, and date of creation.

Since then, four human annotators have examined the selected games. Except for games with poor reviews and games with poor quality reviews, the game was chosen as games with experience in play for annotators' smooth evaluation. Finally, about 5,000 datasets were selected from five games. These reviews were additionally generated by Human Annotators to store reference summaries as pairs.

3.1 Review collection site

Among the various comprehensive review sites currently available, we collected user review information from Metacritic. This website offers high reliability in terms of video game-related data, and its metascore evaluation index has a considerable influence on game evaluation. As of March 25, 2020, the site aggregated reviews for a total of 5,033 games. We collected all reviews—a total of 278,849 reviews—and each item of the collected content comprised the game title being reviewed, author ID, review content, user score, and date of creation.

3.2 Selection of games

Specific games were selected from all games that were reviewed, and the datasets were built as in Table 1. The screening criteria are as follows.

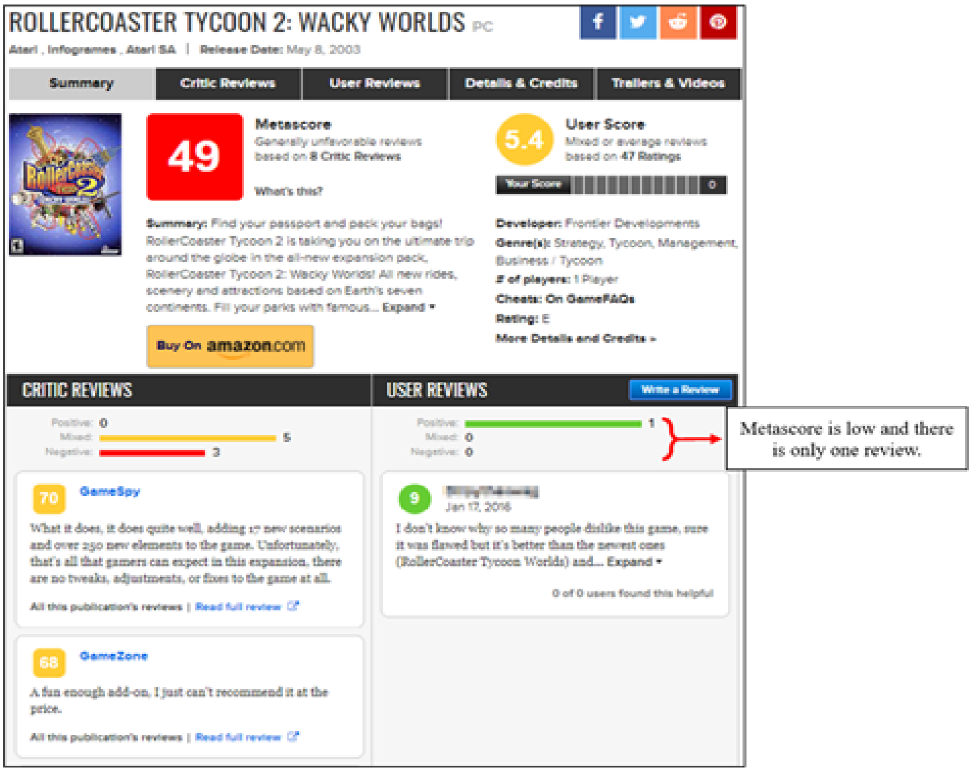
1. Metascore of 50 or higher, avoiding games with significantly low scores.
2. More than 1,000 reviews available, excluding games with fewer reviews.
3. Quality of the game reviewed is above a certain level. To ensure that games are limited to a certain level of quality, 100 reviews per game must be reviewed manually.
4. Limited to games with knowledgeable reviewers. To ensure that actual game quality was considered, only reviews written by reviewers with an experience of playing the actual game were considered.

Table 1. Selected game lists

Game's title	Meta score	Total number of reviews
Portal 2	95	1861
BioShock infinite	94	1586
Overwatch	91	1231
Borderlands 3	81	1085
Witcher 3	93	1709

If the game failed the first criteria, the low recognition and perceived quality resulted in a low number of user reviews. Fig. 2 shows one of examples for that. Even for highly recognized games as shown in Fig. 1, most were unilaterally criticized while the rest of the reviews included a high amount of slang. Thus, we excluded these games from the dataset screening process. The remaining games were then checked for the overall quality of the reviews before analysis. For example, Fig. 3 show low-quality reviews and high-quality reviews, respectively. Through the quality check process, low-quality reviews were filtered out. Next, the minimum number of reviews required to proceed with the selection was believed to be about 1,000, and therefore, only games with more than 1,000 reviews were selected.

For example, as in Fig. 1, because it is difficult to obtain good summarization results when there are many slang words, special characters, or meaningless words in the dataset, the third criterion was used to investigate the proportion of irrelevant data. After randomly selecting 100 reviews for each game, the quality of the reviews was examined; games that did not pass the examiner's evaluation were removed. Because writing a relevant game review requires knowledge of the target game, the final criteria limits game selection to those with reviewers that had real play experience.



ROLLERCOASTER TYCOON 2: WACKY WORLDS PC

Atari , Infogrames , Atari SA | Release Date: May 8, 2003

Metascore
Generally unfavorable reviews based on 8 Critic Reviews

What's this?

User Score
Mixed or average reviews based on 47 Ratings

Summary: Find your passport and pack your bags! RollerCoaster Tycoon 2 is taking you on the ultimate trip around the globe in the all-new expansion pack, RollerCoaster Tycoon 2: Wacky Worlds! All new rides, scenery and attractions based on Earth's seven continents. Fill your parks with famous... Expand ▾

Developer: Frontier Developments
Genre(s): Strategy, Tycoon, Management, Business / Tycoon
# of players: 1 Player
Cheats: On GameFAQs
Rating: E
More Details and Credits »

CRITIC REVIEWS

Positive: 0
Mixed: 5
Negative: 3

70 GameSpy
What it does, it does quite well, adding 17 new scenarios and over 250 new elements to the game. Unfortunately, that's all that gamers can expect in this expansion, there are no tweaks, adjustments, or fixes to the game at all.
All this publication's reviews | Read full review

68 GameZone
A fun enough add-on, I just can't recommend it at the price.
All this publication's reviews | Read full review

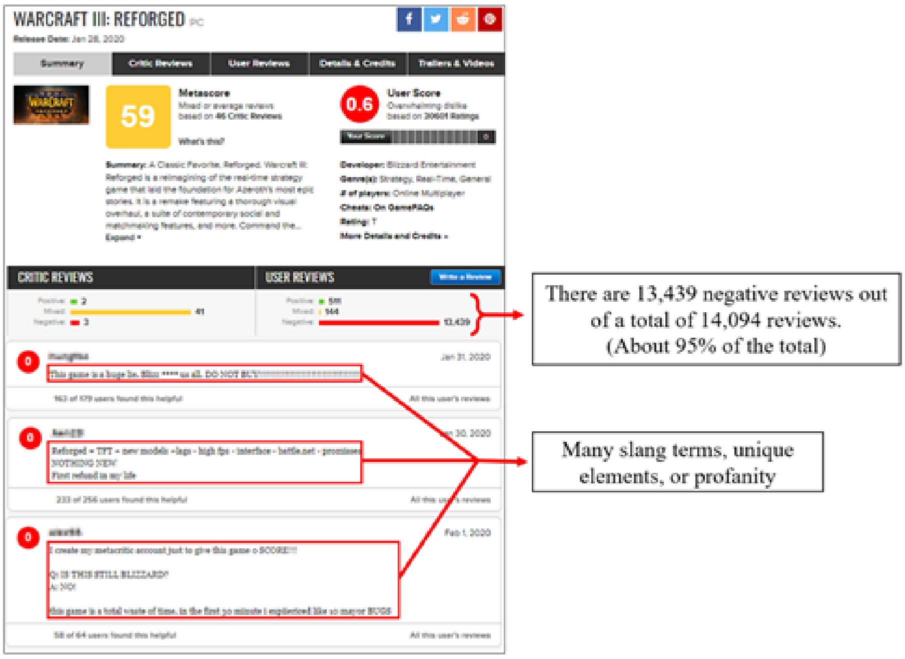
USER REVIEWS

Positive: 1
Mixed: 0
Negative: 0

9 [redacted]
Jan 17, 2016
I don't know why so many people dislike this game, sure it was flawed but it's better than the newest ones (RollerCoaster Tycoon Worlds) and... Expand ▾
0 of 0 users found this helpful

Metascore is low and there is only one review.

Fig. 1. Game with many reviews but with reviews that include many slang terms, unique elements, or profanity – Warcraft 3: Reforged



WARCRAFT III: REFORGED PC

Release Date: Jan 28, 2020

Metascore
Mixed or average reviews based on 46 Critic Reviews

What's this?

User Score
Overwhelming dislike based on 30601 Ratings

Summary: A Classic Favorite, Reforged. Warcraft III: Reforged is a reimagining of the real-time strategy game that laid the foundation for Azeroth's most epic stories. It is a remake featuring a thorough visual overhaul, a suite of contemporary social and matchmaking features, and more. Command the... Expand ▾

Developer: Blizzard Entertainment
Genre(s): Strategy, Real-Time, General
# of players: Online Multiplayer
Cheats: On GameFAQs
Rating: T
More Details and Credits »

CRITIC REVIEWS

Positive: 2
Mixed: 41
Negative: 3

USER REVIEWS

Positive: 511
Mixed: 144
Negative: 13,439

0 [redacted] Jan 31, 2020
This game is a huge lie. Bliz **** us all. DO NOT BUY!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
163 of 179 users found this helpful | All this user's reviews

0 [redacted] Jan 30, 2020
Reforged = TFT + new models +lags - high fps - interface - battle.net - promises
NOTHING NEW
First refund in my life
233 of 256 users found this helpful | All this user's reviews

0 [redacted] Feb 1, 2020
I create my metacritic account just to give this game 0 SCORE!!!

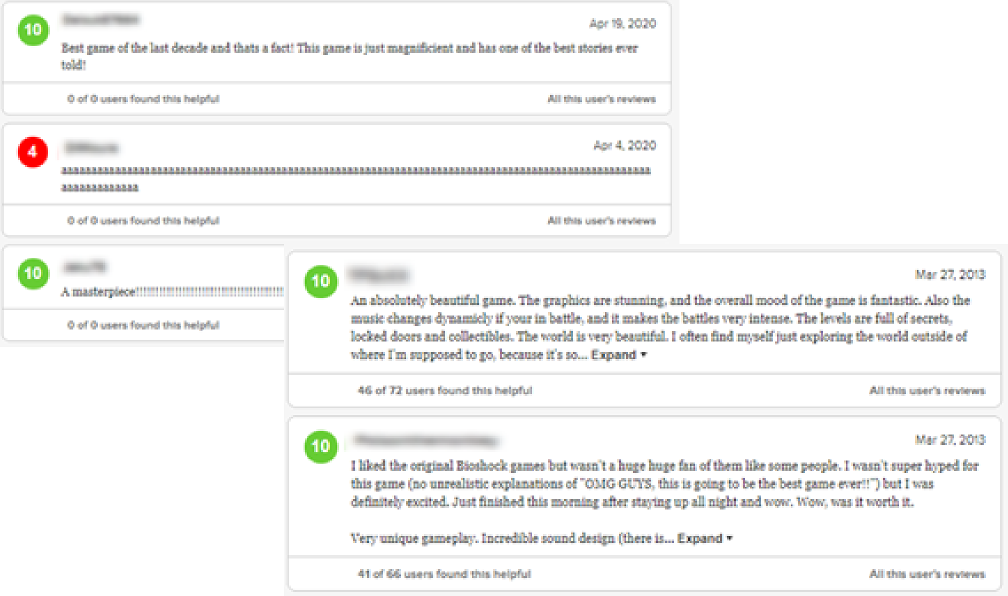
Q: IS THIS STILL BLIZZARD?
A: NO!

this game is a total waste of time. in the first 30 minute i explerfied like 10 mayor BUGS
58 of 64 users found this helpful | All this user's reviews

There are 13,439 negative reviews out of a total of 14,094 reviews. (About 95% of the total)

Many slang terms, unique elements, or profanity

Fig. 2. Example games with a score of less than 50 points and fewer reviews – Rollercoaster Tycoon 2: Wacky Worlds



10 [redacted] Apr 19, 2020

Best game of the last decade and thats a fact! This game is just magnificient and has one of the best stories ever told!

0 of 0 users found this helpful All this user's reviews

4 [redacted] Apr 4, 2020

aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

0 of 0 users found this helpful All this user's reviews

10 [redacted]

A masterpiece!!!

0 of 0 users found this helpful

10 [redacted] Mar 27, 2013

An absolutely beautiful game. The graphics are stunning, and the overall mood of the game is fantastic. Also the music changes dynamicly if your in battle, and it makes the battles very intense. The levels are full of secrets, locked doors and collectibles. The world is very beautiful. I often find myself just exploring the world outside of where I'm supposed to go, because it's so... Expand ▾

46 of 72 users found this helpful All this user's reviews

10 [redacted] Mar 27, 2013

I liked the original Bioshock games but wasn't a huge huge fan of them like some people. I wasn't super hyped for this game (no unrealistic explanations of "OMG GUYS, this is going to be the best game ever!!") but I was definitely excited. Just finished this morning after staying up all night and wow. Wow, was it worth it.

Very unique gameplay. Incredible sound design (there is... Expand ▾

41 of 66 users found this helpful All this user's reviews

Fig. 3. Examples of low-quality reviews and high-quality reviews

3.3 Review analysis for the selected games

About 1,000 reviews were re-selected for each of the 5 games selected through the process discussed in Section B. The reason for selecting around 1,000 data points is that it is unreasonable to inspect a larger amount of data manually, and if too few data points are selected, the amount of data in the dataset is too small. Therefore, we limited the numbers to approximately 1000±100. In the selection process, non-English reviews were excluded, and reviews were selected based on the number of user endorsements received. The reason is that the fewer are the number of user endorsements, the higher is the amount of profanity or meaningless texts that is part of the reviews, which makes it difficult to use them as valid review data.

3.4 Determining whether a summary is present

Reviews containing summary sentences were selected to use them as training data in machine learning-based game review summarization. Selection processes were conducted manually by four annotators who checked the criteria for each game review and its summary based on the conditions provided below. The annotators reviewed each document by cross-verifying it against these criteria.

1. Whether to re-refer the content in the sentence as mentioned above
2. If a sentence contains some of the assessment items
3. If the writer's opinion is included in the sentence
4. Exclude long sentences that cannot be called a summary

a) Check for Rule 1: The summary was determined from the perspective of recall, and the assessment was possible from the same perspective when comparing with the Recall- based ROUGE.

b) Check for Rule 2: The assessment items mentioned in Section III.C were used; reviews that could not be assessed by relatively objective evaluation were excluded.

c) Check for Rule 3: The author determined from the review that the key content he wants to convey is contained in that sentence, which is the basis for the summary's fundamental role.

d) Check for Rule 4: Criteria were applied to exclude sentences in exceptional cases such as when the text was too short or when the text was about 1,000 characters long.

Further, the analysis was performed to add a summary to the criteria where a particular linking phrase appeared in the process; however, this was excluded because it was not suitable as an absolute indicator.

3.5 Statistic of game review dataset (After refinements)

In the final version, a total of five games were selected and utilized first; about 1,000 reviews were evaluated item-by-item with the criteria provided in Section 3.4.

Among these approximately 1,000 reviews, a total of 400 summaries were selected following the filtering procedure mentioned in Section 3.2. Table 2 shows the statistics of our game review dataset.

Table 2. Dataset configuration after refinements

Game's title	Meta score	Total number of reviews	Summary
Portal 2	95	1861	137
BioShock infinite	94	1586	86
Overwatch	91	1231	28
Borderlands 3	81	1085	56
Witcher 3	93	1709	93
Sum	☒	5220	400

Ⅳ. Experiments

4.1 Compared models and settings

We conducted game review summarization using state-of-the-art pre-trained models (i.e., BERT and RoBERTa). The collected and refined game review dataset in Section 3 was used for the experiments. The dataset was randomly divided into the training set and testing set in a ratio of 8:2. Using the training set, the pre-trained models were fine-tuned as our game review summarization models. The summarization results were evaluated on the test set using the ROUGE score. There are several types of ROUGE scores. In this paper, ROUGE-1, ROUGE-2, and ROUGE-L, which set scores according to the match between reference summaries and generated summaries for one (Unigram), two (Bigram), and N (N-gram), were used.

For a set of experiments, we performed a fine-tuning through transfer learning using our game review dataset based on the pre-trained model, BERT provided by Google.

In case of the basic BERT model, the BERT-large-cased model and the transformer model provided by HuggingFace were used as extractive and abstractive models, respectively. For the BertSum model, a pre-trained model was used as a combined extractive and abstractive model. Finally, the MatchSum model was used to extract summaries by matching the contextual representations of the document with the summaries by utilizing the RoBERTa model, which is a BERT model pre-trained with CNN/ DailyMail datasets .

4.2 Compared models and settings

The experimental results listed in Table 3 confirm that the overall score was lower than the existing state-of-the-art score. This result occurred because of the differences in the dataset. Most existing state-of-the-art results used standardized news data such as that from CNN or the Daily Mail. Therefore, relatively free-form reviews such as game reviews showed low scores.

Table 3. Experiment result

Model	ROUGE-1	ROUGE-2	ROUGE-L
BERT(Transformer)-Extract	34.01	25.44	33.91
BERT(Transformer)-Abstract	26.65	15.30	25.90
BertSum(ExtAbs)	24.89	13.26	23.17
MatchSum(RoBERTa)	27.02	10.15	22.31

In the extractive summarization method, the ROUGE score was calculated to be relatively high; this was attributed to the examiners using a method similar to the extractive method for generating the correct summary. In the case of MatchSum using RoBERTa, and this showed a score in the mid-20s, which is similar to the current result values.

Another reason besides the characteristics of the dataset is that the amount of game review data used for fine-tuning is small. In terms of fine tuning, the higher is the similarity of the data, the more effective are results that can be derived even with a small amount of data. However, we used game reviews having very different characteristics from general newspaper data, and the amount of data used for learning is small; these reasons may have contributed to a disparate result compared to those of the state-of-the-art models.

Most of the ROUGE-2 scores were low except for Transformer-Extract (BERT). Because the original text was not well used in the summary due to the characteristics of Abstractive Summarization. And also because the summary of the dataset that we built was long.

4.3 Case study with summary samples

In the case of BERT (Transformer)-Extract, which had the highest average ROUGE value among the results, a high-quality summary sentence corresponding to the high score was generated as shown below in the result for (a) in Table 4.

However, there are frequent cases where the score is reduced because of the presence of longer sentences. (b) illustrates a representative example that includes the keyword “Amazing,” and if you read the entire contents of the extracted sentence, it is generally favorable for the game. However, the actual ROUGE score was evaluated for a summary with a low score of 11.11 based on ROUGE-1.

Among the abstractive techniques, the state-of-the-art BertSum-based model－BertSumExt Abs－showed a questionable result. In the case of (c), a summary of low quality was generated with a low score.

Table 4. Analysis results in sentence units

	Generated summary	Reference summary	ROUGE-1
	BERT(Transformer)-Extract		
(a)	Portal 2 is definitely Valve’s best game up to date! The game is absolutely brilliant. The level design is a lot better than in portal 1.	The game is absolutely brilliant. The level design is a lot better than in portal 1. The graphics are great and the gameplay is just superb.	58.82
(b)	This game is amazing. I usually just cling onto Online multiplayer games, but i’ve literally been playing this game past week none-stop. It’s very well polished.	Amazing story, Amazing graphics. Don’t loose out on this game.	11.11
	BertSum(ExtAbs)		
(c)	The new gel and laser are really fun to play with the co-op mode is super fun	The best game ever (Valve and not). Gameplay, graphic, soundtrack, story every single thing of this game it’s simply wonderful.	5.4
(d)	The best narrative in video game history is the best in the video game brilliant voice-controlled voice acting and best overall narrative in history	The engaging, mind-bending gameplay of the original with a few twists, the same fantastic voice acting, and the best narrative in video game history. Rarely is the medium handled with such technical and artistic proficiency.	41.37

However, in the case of (d), i.e., in the case of the score, it is a low- quality summary sentence that repeats the same content by mentioning “voice” along with key points and frequently using the word “best.” This summary received a high score of 41.37.

As a result, in the case of the ROUGE score as shown in Table 4, when looking at each case, the quality of the generation summary was better, but the score was lower. And there were cases where the quality was bad but the score was high. As such, it was confirmed that there was a difference between the ROUGE score and the actual result, and through Case Study, the difference was compared using human evaluation in Section 5.

Ⅴ. Employing Game Review Factors

5.1 Five Quality Factors (QFs)

Like movies’ evaluation elements [13], as the production value of cut scenes in games is increasing, graphics and sound have started to weigh in on the gameplay experience, and therefore, it appears more consistently in reviews

In addition, the graphics, error, and system criteria reflect the game's other characteristics. For graphics, the visual quality of games has developed sufficiently to be evaluated as art, and related exhibitions have been held [14]. During our evaluation, we observed that graphical and visual elements were mentioned frequently in the reviews, and they were thus added to our evaluation criteria.

The Error criterion is a prerequisite for playing the game smoothly. Therefore, to describe user experiences, it is essential to determine whether it is possible to play the game smoothly. The system criterion was added to evaluate special systems in the game and the technical optimization of the game.

5.2 Rethinking ROUGE with the QFs

To find out existing gaps and similarities between ROUGE scores and our proposed QFs, we analyzed five different sample summaries selected from five different games.

For the analysis, QF scores for each of the original text, reference summary, and generated summary were scored through 4 human annotators. Five games in the data set were evaluated, and Table 5~9 show the results.

In the case of BioShock Infinity (Table 5), because the story was very important, there were many mentions about the game’s story itself.

Table 5. (a) Bioshock: Infinite review example – result of BertSum

Original text	Though not as gritty as the previous two installments, Bioshock Infinate still boasts a brilliant cohesive storyline with well drawn-out characters and a setting that is visually impressive. The graphics are great, the voice acting is great, the ending is convoluted and confusing but its the score that helped me get through the game. The score is none-short of amazing. Overall the game isn't as good as the previous two but it's still a fantastic game with vast improvements and a great AI system where the AI actually help you out and not get in the way of thing.
Reference summary	Overall the game isn't as good as the previous two but it's still a fantastic game with vast improvements and a great AI system where the AI actually help you out and not get in the way of thing.
Generated summary	Though not as gritty as the previous two installments, bioshock infinate still boasts a brilliant cohesive storyline with well drawn-out characters and a setting that is visually impressive. the score is none-short of amazing

Table 5. (b) Bioshock: Infinite review evaluation by human annotator and ROUGE score

	Graphic	Stroy	Sound	Error	System	Sum
Original text	2	2.5	2.5	1	3.5	11.5
Reference summary	1	1	1	1	3.5	7.5
Generated summary	1.5	2	1.5	1	1	7
	ROUGE-1		ROUGE-2		ROUGE-L	
	32.88		11.27		29.03	

Table 6. (a) Borderlands 3 review example – result of BertSum

Original text	Even though Randy is a pedophile, liar and a terrible human being overall, I gotta admit he is capable of making good games (even though I'm not sure if he engages in the proccess itself or just supervisor). …… While the story, humour, characters feel choped in half in quality and downgraded in scenery, the game still provides you a good shooter experience and overall feels like a good time spent.
Reference summary	Conclusion: Borderlands 3 keeps and refines its brilliant gunplay, reinforces class progression and design and enhances visual art style and presents you good amount of fun content and dozens guns and gear to loot, just like a good looter-shooter supposed to do. While the story, humour, characters feel choped in half in quality and downgraded in scenery, the game still provides you a good shooter experience and overall feels like a good time spent.
Generated summary	Even though randy is a pedophile and a terrible human being overall, he is capable of making good games.the good: the gunplay is solid, class progression feels satisfying and fun. visual design is still gorgeous and some characters are nicely written

Table 6. (b) Borderlands 3 review evaluation by human annotator and ROUGE score

	Graphic	Stroy	Sound	Error	System	Sum
Original text	2.5	4	1	3	4.5	15
Reference summary	3	2.5	1	1	4	11.5
Generated summary	1.5	1.5	1	1	3.5	8.5
	ROUGE-1		ROUGE-2		ROUGE-L	
	25.86		1.75		13.18	

Table 7. (a) Portal 2 review example – result of BertSum

Original text	The singleplayer in Portal 2 was like its predecessor, a very well written story, hilarious moments, and tricky test chambers. The environment has changed a bit, and new features has been added to the chambers, for example lasers. If you're decent at solving the puzzles, you'll complete the game in a matter of hours. What I personally have not tried is the co-op mode, but from what I take it enriches the game even more. I only detected a few minor glitches, mostly graphical, but that was a rarity. I believe Portal 2 is one of the most polished games I've tried this year. I recommend buying it, especially if you have played Portal 1. I'm giving it a 10 out of 10 for the story, gameplay, graphics, music and sound..
Reference summary	I'm giving it a 10 out of 10 for the story, gameplay, graphics, music and sound.
Generated summary	The singleplayer in portal 2 was like its predecessor. a very well written story, hilarious moments and tricky test chambers. new features has been added to the chambers

Table 7. (b) Portal 2 review evaluation by human annotator and ROUGE score

<table>
<tr><th></th><th>Graphic</th><th>Stroy</th><th>Sound</th><th>Error</th><th>System</th><th>Sum</th></tr>
<tr><td>Original text</td><td>2</td><td>2.5</td><td>2</td><td>2.5</td><td>3</td><td>12</td></tr>
<tr><td>Reference summary</td><td>2</td><td>2</td><td>2</td><td>1</td><td>2</td><td>8</td></tr>
<tr><td>Generated summary</td><td>1</td><td>1</td><td>1</td><td>1</td><td>3</td><td>7</td></tr>
<tr><td rowspan="2"></td><td colspan="2">ROUGE-1</td><td colspan="2">ROUGE-2</td><td colspan="2">ROUGE-L</td></tr>
<tr><td colspan="2">9.3</td><td colspan="2">0</td><td colspan="2">9.75</td></tr>
</table>

Table 8. (a) Overwatch review example – result of BertSum

Original text	Overwatch is a game which has a good concept but a bad implementation. Poor map and mechanics design, poor graphics, a serious lack of writing talent, a featureless matchmaking system, and the microtransactions added on top are all factors which have contributed to Overwatch's failure to provide a fun experience. With a single sentence, the closest one can get to describing the massive ……. In short, Overwatch is a good game concept with bad game design. Arbitrary mechanic design, poor map design, poor writing, well-performing yet poor quality graphics, a barely functional matchmaking system that punishes the player for its failures, and microtransactions as a rotten cherry on top of an ant-ridden ice cream sundae make it clear to label this game with the bottom line of: Avoid.
Reference summary	In short, Overwatch is a good game concept with bad game design. Arbitrary mechanic design, poor map design, poor writing, well-performing yet poor quality graphics, a barely functional matchmaking system that punishes the player for its failures, and microtransactions is terrible.
Generated summary	Overwatch is a game which has a good concept but a bad implementation. the base design of the game is both unique and functional, but where the developers go wrong is in the specifics

Table 8. (b) Overwatch review evaluation by human annotator and ROUGE score

<table>
<tr><th></th><th>Graphic</th><th>Stroy</th><th>Sound</th><th>Error</th><th>System</th><th>Sum</th></tr>
<tr><td>Original text</td><td>4</td><td>3</td><td>1</td><td>1</td><td>4</td><td>13</td></tr>
<tr><td>Reference summary</td><td>2</td><td>2</td><td>1</td><td>1</td><td>4</td><td>10</td></tr>
<tr><td>Generated summary</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td><td>8</td></tr>
<tr><td rowspan="2"></td><td colspan="2">ROUGE-1</td><td colspan="2">ROUGE-2</td><td colspan="2">ROUGE-L</td></tr>
<tr><td colspan="2">32.43</td><td colspan="2">5.55</td><td colspan="2">26.67</td></tr>
</table>

Table 9. (a) Witcher 3 review example – result of BertSum

Original text	The game is awesome! From the start there is so much to do. Animations for everything have been improved drastically and the combat feels more fluid than witcher 2. The story is great as always, every little sidequest has a story behind it no fetch quest level stuff. …… So overall I will take away a point for the graphics/performance being not as good as everyone hoped. TLDR: the gameplay is intact, controls are good, story is good, sidequests are non fetch quests, performance is ok for some, not great for others, graphics are good at times other areas a little disappointing but not enough to deter from gameplay.
Reference summary	The gameplay is intact, controls are good, story is good, sidequests are non fetch quests, performance is ok for some, not great for others, graphics are good at times other areas a little disappointing but not enough to deter from gameplay.
Generated summary	The game is awesome! from the start there is so much to do<q>animations for everything have been improved drastically and the combat feels more fluid than witcher 2<q>every little sidequest has a story behind it no fetch quest level stuff

Table 9. (b) Witcher 3 review evaluation by human annotator and ROUGE score

	Graphic	Stroy	Sound	Error	System	Sum
Original text	4	3	1	1	4.5	13.5
Reference summary	2	2	1	1	4	10
Generated summary	1	1	1	1	3.5	7.5
	ROUGE-1		ROUGE-2		ROUGE-L	
	22.22		0		8.57	

The personalities of the game characters were excellent and the game itself was systematically excellent; therefore, there were many cases where system items were described in detail. Reflecting this fact, both the original text and the reference summary received a high score for the system item among the five QFs, but the generated summary received a low score. This indicates that the generated summary did not well summarize what was mainly mentioned in the review. Unlike these results, the ROUGE score received a high score. Other games also showed different aspects of QFs evaluation and ROUGE score. Through this, we confirmed once again that even a summary with a high ROUGE score could not be convinced in the field of game review.

Ⅵ. Conclusion

ROUGE is mainly used in the fields of text generation and text summarization. In this study, we identified that ROUGE metric is not feasible for game review summarization any more. To this end, we showed the limitations of ROUGE scores by measuring the performance of automatically generated review summaries using SOTA algorithms. In addition, we perform a comparative analysis using game review QFs with five different game reviews.

As our future work, we plan to develop an automatic text summarization technique that considers the QFs because it revealed promising pieces of evidence that the limitations of the ROUGE score can be mitigated by computing QF scores.

References

[1] T. Wijman, "The World’s 2.7 Billion Gamers Will Spend $159.3 Billion on Games in 2020", Newzoo.com, 2020. https://newzoo.com/insights/articles/newzoo-games-market-numbers-revenues-and-audience-2020-2023/ [accessed: Dec. 09, 2020].

[2] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", Proc. Work. text Summ. branches out (WAS 2004), No. 1, pp. 25–26, 2004, [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85.

[3] N. Schluter, "The limits of automatic summarisation according to rouge", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Valencia, Spain, Vol. 2, Short Papers, pp. 41–45, Apr. 2017.

[4] W. Tay, A. Joshi, X. Zhang, S. Karimi, and S. Wan, "Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation", in Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, Sydney, Australia, pp. 52-60, Dec. 2019, [Online]. Available: https://www.aclweb.org/anthology/U19-1008.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., Minneapolis, Minnesota, Vol. 1, pp. 4171–4186, Jun. 2019.

[Online]. Available: http://arxiv.org/abs/1810.04805.

[6] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive Summarization as Text Matching", in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, [Online], pp. 6197-6208, Jul. 2020, https://doi.org/10.18653/v1/2020.acl-main.552.

[7] Y. Liu, "Fine-tune BERT for Extractive Summarization", arXiv, Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.10318.

[8] D. Antognini and B. Faltings, "GameWikiSum: a Novel Large Multi-Document Summarization Dataset", arXiv, Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.06851.

[9] S. Pecar, "Towards opinion summarization of customer reviews", in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop, Melbourne, Australia, pp. 1-8, Jul. 2018, https://doi.org/10.18653/v1/p18-3001.

[10] G. Panagiotopoulos and G. Giannakopoulos, "A Study on Video Game Review Summarization", in Proceedings of the Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources associated with RANLP 2019, Varna, Bulgaria, pp. 35-43, Sep. 2019, https://doi.org/10.26615/978-954-452-058-8_006.

[11] K. Yauris and M. L. Khodra, "Aspect-based summarization for game review using double propagation", in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), Denpasar, Indonesia, pp. 1-6, Aug. 2017, https://doi.org/10.1109/ICAICTA.2017.8090997.

[12] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv, no. 1, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[13] F. M. Schneider, "Measuring Subjective Movie Evaluation Criteria: Conceptual Foundation, Construction, and Validation of the SMEC Scales", Commun. Methods Meas., Vol. 11, No. 1, pp. 49-75, Jan. 2017, https://doi.org/10.1080/19312458.2016.1271115.

[14] C. Varnava, "The art of video games", Nature Electronics, 2018. https://americanart.si.edu/exhibitions/games [accessed: Nov. 22, 2020].

Authors

Yeong-Hun Lee

2020 ~ 2022 : MS degree in Department of Computer Engineering Kumoh National Institute of Technology (KIT)

2020 : BS degree in Department of Computer Engineering Kumoh National Institute of Technology (KIT)

Research interests : Text Summarization, Abstractive Summarization

Yuchul Jung

2005 ~ 2011 : PhD degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST)

2009 ~ 2013 : Senior Researcher at Telecommunications Research Institute (ETRI)

2013 ~ 2017 : Senior Researcher at Korea Institute of Science and Technology Information (KISTI)

2017 ~ 2022 : Assistant professor in Department of Computer Engineering, Kumoh National Institute of Technology (KIT)

2022 ~ present : Assistant professor in Department of Artificial Intelligence Engineering, Kumoh National Institute of Technology (KIT)

Research interests : Machine learning based NLP (text mining, sentiment analysis, automatic knowledge base construction, etc.), Korean speech recognition, and Medicine 2.0.