Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Manga Scene Estimation by Quiz Question and Answer

Tsubasa Sakurai[a], Yume Tanaka[a], Yuto Sekiguchi[a], Satoshi Nakamura[a]

[a]*Meiji University, Nakano 4-21-1, Nakano-ku, Tokyo, Japan*

Abstract

In reading manga, it is common to look back at the storyline when following a serialized work. Although there are services that assist comic re-reading through quizzes, searching for specific parts related to the quiz takes a lot of time and complicates the review process. Therefore, in this study, we examined whether it is possible to estimate the scenes related to the quiz based on the quiz questions, answers, and manga-specific features. To achieve this, we extracted key elements from the comic and proposed two estimation methods: a word-based CS method and a context-based GPT method. Furthermore, we discussed extractable and difficult-to-estimate scenes in comics. The results showed that the pages containing the answers could be estimated with a probability of 66.7%. Pages containing specific keywords or events were easier to estimate, while those requiring an understanding of the comic's overall time series and context were more difficult to estimate. In addition, since the accuracy varied greatly depending on the presence or absence of the answer text, it can be considered that the content as close as possible to the topic of the quiz can be estimated if important keywords such as the answer text are included.

Keywords: Comic; Quiz; Scene Estimation; GPT-4;

1. Introduction

In Japanese serials, comics are updated on a weekly or monthly basis, and some works continue to be serialized for hundreds of episodes or dozens of volumes. Therefore, it is difficult to find a specific scene when we want to look back at the previous episodes. Although some of the past stories are summarized on the Internet or SNS, you may face spoilers for parts of the story that you have not yet read. Some studies have tried to improve the understanding of stories and characters by automatically constructing character relationship diagrams [1][2]. Although these can be used to look back and improve understanding of the content, they also increase the burden on the reader when reading.

Thus, we proposed a method called "ComiQA" (https://comiqa.com) that enables users to remember the content of a comic by creating quizzes and answering them, and they can share the quizzes with others. Before the user reads the newest volume of a comic, he/she can easily recollect the content of its previous volume by trying to answer these quizzes. Currently, there are a total of 1,460 quizzes from 696 comics registered (as of April 25, 2024). If the user cannot answer a question on a quiz, he/she can re-read that part of the comic. However, it is not easy for them to find the specific page. This problem occurs not only when answering questions on ComiQA, but also when trying to recall or locate scenes you want to remember or refer to within the comic.

To solve this problem, we propose a method for finding pages from quiz questions and answers. We investigate a method used to help in reading and re-reading a comic by finding relevant scenes from the questions and answers of a quiz on the comic using “ComiQA,” while also considering the elements of comics.

2. Related Work

To utilize the elements of comics, it is necessary to extract the elements from comic images and make use of them. This section refers to research on the technical aspects of comic elements and research that is being carried out on their utilization.

There are various studies that attempt to extract elements in comics, such as panels and lines, using image recognition. Nguyen et al. [4] reviewed the definition of panels in comics and proposed a method for extracting panels. Chu et al. [5] proposed a method for character face recognition, and Tolle et al. [6] proposed a method for highly accurate recognition of lines. In these studies, recognition was performed using comic images and information on text, and it can be said that simpler and more accurate methods are being established. In addition, as a study utilizing elements of comics, Chen et al. [7] proposed an algorithm for understanding multilingual 4-frame comics. Park et al. [8] analyzed the characteristics of comic characters to realize a comic retrieval system using comic characters. Nguyen et al. [9] proposed a Comic MTL model that learns multiple tasks and analyzes the relationship between speech balloons and lines. These studies are necessary to develop systems that take into account the content of comics, and one of the methods for understanding comics is to use LLM. Vivoli et al. [10] introduced the Multimodal-LLM architecture for the Text-Cloze Task and achieved almost the same accuracy as BLIP with one-fifth of the parameters by using adaptive learning. Guo et al. [11] also proposed a Multimodal Manga Complement (M2C) task that combines visual features and text, showing that it can effectively solve problems such as missing pages and text degradation.

Research is conducted to support understanding and reflection on the content of a work. For instance, studies include the automatic construction of character relationship diagrams for comics [1] and the creation of character networks in literary texts [2], and Matsui et al [3] proposed an image retrieval system for comics using sketches. These studies have enabled visual networks to deepen understanding of the work’s content and retrieve the content from vague information. However, research and services on methods to easily reflect on comic stories or specific scenes have not yet been fully researched.

Considering these factors, the present study focuses on quizzes on comics. Quizzes on “ComiQA” are composed of keywords and vague sentences. Also, it is possible to construct a dataset in a format suitable for estimation, in which pages with the correct answers can be registered. The estimation of comic scenes from keywords or limited information for contents consisting of images and text is a theme close to the field of video-scene retrieval. Various studies have been conducted on video-scene retrieval. Zachariah et al. [12] proposed a video retrieval system QIK+ that effectively captures scene order and object relationships using natural language processing and scene captions, and showed high accuracy on the MSR-VTT dataset. Baraldi et al. [13] proposed a video retrieval system that divides videos into coherent scenes and provides visually appealing thumbnails for queries. Qi et al. [14] also proposed an online cross-modal scene retrieval framework for image and text data, optimized for streaming data. However, few studies have been conducted on content such as comics. Therefore, this study estimates the pages containing the answers to the quiz on comics. We then contribute to research on retrieving comic scenes from vague information and supporting comic recall.

3. Quiz Dataset

In ComiQA, there are many quizzes, but there is a tendency for many of the quizzes to be created for the latter half of a volume. In addition, the number of quizzes is not enough to check the accuracy of estimating certain pages. For this reason, we recruited collaborators to increase the number of quizzes on pages of comic books for evaluation experiments with ComiQA. Then, we used the quizzes' questions, answers, and related pages in ComiQA as a quiz dataset.

The quizzes to be used in the experiment should not require any prerequisite knowledge, and the content should be easily recognizable from within the volume. In addition, quizzes that are too simple in content or have little relevance to an episode (e.g., "What is the full name of Q. ___?") are not appropriate for estimating the corresponding page. Therefore, it is preferable that the target comics can be read without prerequisite knowledge and are relatively complex in content.

Based on the above, we selected sports-related comics because of the large number of characters and because they are often set in schools. The target comics were limited to the first volume, as this would not require any prerequisite knowledge, and had at least nine characters appearing in the first volume. The selected sports-related comics are as follows.

- Farewell, My Dear Cramer
- Aoashi
- Haikyu!!
- Asahinagu
- The Prince of Tennis
- Sokyu Boys
- Tokyo Toy Boxies
- Days

We asked the collaborators to create three quizzes for each of the eight comics, one quiz each from the beginning, middle, and end of the story (three quizzes in total), based on the information in the volume, and told them that they were free to create the quizzes (question, answer, and its page number) while looking back at the comic.

Five collaborators, three men and two women, who read comics on a daily basis were selected to participate in the construction of the dataset. They registered three quizzes for each of the eight comics, resulting in a total of 120 quizzes. After that, we modified the page number if the registered page number was mistaken. In this study, a total of 138 quizzes, including already created quizzes, were used to estimate the corresponding page.

Fig. 1 shows the number of quiz pages and the number of quizzes created for each comic. The horizontal axis shows the percentage of the total number of pages in the comics. The results show that quizzes were created from all page sections except for the very beginning, indicating a relatively even distribution of quizzes from the beginning to the end of the comics.

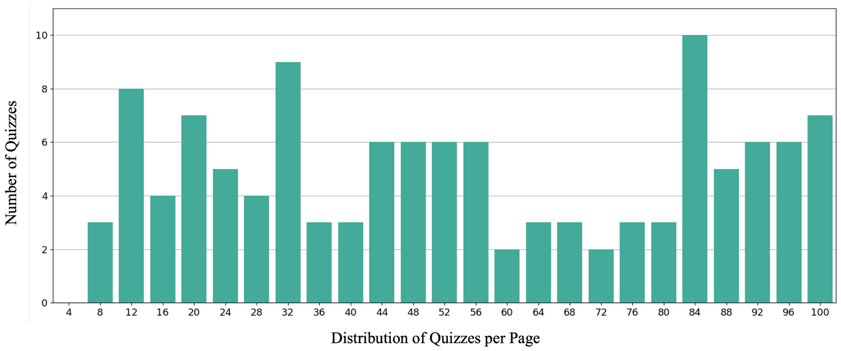


Fig. 1. Number of quizzes created per page
(the horizontal axis shows where the quizzes were created as a percentage of the total pages).

4. Estimation from Quiz Using Comic Elements

4.1. Selection of Comic Elements

Comics are composed of text and images, and the way that they are depicted varies greatly depending on the genre and characteristics of the work. Therefore, it is important to obtain information from both the text and images. In addition, when creating a quiz, it is important to answer questions related to "What," such as "What did ___ do here?" and "Who did ___?", and in both cases, the information about the characters and their actions is considered to be a clue.

On the basis of these considerations, we treat information from the lines, images, and characters in the comic as particularly important elements in finding the answers to quiz questions. Specifically, line information refers to all textual information in the comic, such as lines in and out of speech balloons and narration. As for information obtained from images, each panel in a comic is treated as an image that represents some action or situation, and the content occurring in the image is treated as text information. Character information includes the names of the characters and where they appear on each page. For each page, the number of times each character appears is extracted, and then the page information and character names are obtained.

4.2. Extraction of Comic Elements

For this paper, we used mokoro [15], which can detect and recognize text information with OCR technology, to extract lines in a comic. In addition, we used BLIP [16], which provides trained models as a VLP framework to extract descriptions from frames. Then, we automatically translated the extracted descriptions from English to Japanese for processing when page estimation was performed in Japanese. Additionally, we used the comic-panel-detectors API of Roboflow [17] to detect the frame area. Here, to ensure accurate frame extraction, we adjusted parameters and manually eliminated some false positives.

We manually annotated the number of appearances of each character. The number of appearances was collected on a page-by-page basis. For the collection, we treated characters that met the following definitions as appearing characters.

- The character talks at least once.
- The character's name (full name, family name, or first name) appears in the work.
- The character is called by name in a line.
- The character's face is clearly depicted in the scene.

4.3. Estimation Method

The quiz questions and answers created in a comic may include various keywords such as events in the story and characters appearing in the story. Therefore, in this study, we propose two methods: the "CS method (Cosine Similarity method)", which is based on word estimation, and the "GPT method", which is based on content understanding.

The "CS method" determined the similarity between sentences by calculating the TFIDF values of words from the quiz sentences and text information obtained from the comic, and then calculating the cosine similarity from the TFIDF values. Specifically, a morphological analysis is performed on the elemental information in the comic, obtained for each page, the similarity between these elements and the quiz question text and its answer is calculated, and the page with the highest cosine similarity among all pages is estimated as the answer. The elements used were as follows.

- Quiz question.
- Quiz answer.
- Line information in the comic (feature-lines).
- Text information with image description for each panel (feature-image).
- Whether or not each character appears on each page (feature-appearing).

Table 1. Prompts used to estimate page for quiz and answer. In this example, quiz is "what is Hinata's weapon that Kageyama is killing by his toss?," answer is "swiftness," and its correct page number is 184.

Prompt	Example of output results
Estimate from which page of the comic the following quiz (Question,Answer) was created. Note that the contents of the comic should refer to the information in JSON format shown below. {explanation of the JSON structure}. If the same character is described more than once, it means that the character appears in multiple frames. In the output, please describe the estimated "page number" and the "reason for selecting" for that page. The following is JSON data.	Estimate page number is "184". Reason for selection: In the question text, the keyword "swiftness" is present, and the content of page 184 includes the phrase "fast attack" and the keyword "swiftness". These are elements that indicate Hinata's speed and may refer to the "weapon" feature in question. They are also consistent with the fact that Kageyama's tosses do not take advantage of Hinata's "swiftness," i.e., they are described as "killing it." Therefore, we presume that this page is the source of information regarding the quiz in question.

Candidates for each page were extracted from the quiz sentences (question + answer), line information, image description sentences, and whether each character appears or not. Morphological analysis was performed after concatenating the quiz sentences and answers, as well as other elements in the comic. Since most of the quiz answers were words only, such as keywords, the weight of the words related to the answers was increased. For character information, the name of the character has been added as text information when the character appears on the page.

The "GPT method" is the same up to the process of calculating the cosine similarity, but differs in the estimation of the answer from the cosine similarities on each page. The method extracts the top five pages with the highest cosine similarity and inputs the comic text elements contained in those pages to the GPT-4 Turbo API [18], which then estimates one answer page based on the context of the text. In summary, the two methods in this study are as follows.

- CS method: A method for estimating the page with the highest cosine similarity value between the quiz sentences and each page of the comic.
- GPT method: A method in which GPT estimates the most likely page from the top five pages with the highest cosine similarity between the quiz sentences and each page of the comic.

The information and prompts entered during estimation are as follows (Table 1). Note that the prompts and output at the time of estimation were in Japanese, and the sentences shown in Table 1 are translations of them into English. It should be noted that the JSON data referenced in the prompt contains the dialogue text, caption generation text and appearing character names extracted in section 4.2, each stored as an object with the number of pages as the key.

4.4. Estimation Results

Table 2 shows the accuracy of estimating the pages on which the quizzes were created. Note that considering that each comic quiz was created based on a single scene across multiple pages, we treated the four pages before and after the answer (two-page spread) as correct even if they were included. Each row shows the accuracy with which the CS method and the GPT method were able to uniquely estimate the answer page. The results show that the GPT estimation resulted in an accuracy of 66.7%. However, in the GPT method, there were cases where the top five pages in cosine similarity did not contain the answer. Therefore, we derived the accuracy that GPT can estimate when limited to the situation where there is an answer among the five choices, which resulted in an accuracy of 83.3%.

Table 3 shows the results for different comic elements. Each row shows the type of elements used, and percentages are shown when all elements are used, when one element is excluded from the estimation, and when only lines are used. It can be seen that the estimation was less accurate when other factors were missing. Table 4 shows the accuracy of the estimated pages based only on the question text, without the answers to the quiz sentences. It can be seen that the answer text in the quiz has a significant impact on the accuracy of the estimation.

Table 2. Results of quiz estimation. Estimation using GPT method is a method that estimates the first candidate from the top five candidate pages of cosine similarity.

Estimation method	Accuracy
CS method	55.1%
GPT method	66.7%

Table 3. Comparison of accuracy among comic elements used for estimation.

Used feature			CS method	GPT method
lines	image	appearing		
✓	✓	✓	55.1%	66.7%
✓	✓		52.8%	60.3%
✓		✓	50.0%	65.3%
✓			53.6%	62.0%

Table 4. Accuracy for situations where only information used for estimation is quiz question text.

Quiz sentences used for estimation	Accuracy
Question and Answer	66.7%
Only question	44.2%

5. Discussion and Prospects

According to Table 2, the GPT method had an estimation accuracy of 66.7%, which is higher than that of the CS method. Also, the 83.3% accuracy of GPT when excluding cases where the answer was not included in the answer choices shows that the judgments made by GPT are quite accurate. This may be due to the fact that the combination of estimation by GPT make it possible to understand contexts that cannot be judged only by cosine similarity. Table 3 shows the estimation accuracy under different conditions for different comic elements using the estimation, and it can be seen that removing other elements decreases the accuracy. Therefore, information from comic lines, images, and characters are all important factors in making inferences from quiz text. However, since the degree of accuracy loss will vary depending on the elements removed, it is necessary to consider how to combine these multimodal pieces of information. Table 4 also shows a comparison of the quiz text input for estimation when both the quiz question and answer are used and when only the question text is used, indicating that the estimation accuracy is very low (44.2%) when no answer is provided. This may be due to the fact that quiz answers often contain important keywords or sentences that explain the situation of the answer page scene.

Next, we provide the characteristics of quizzes that can and cannot be estimated and discuss them with examples. For quiz estimation, it is important to consider the entire story as a scene that was related to the quiz other than the answer page. Therefore, Table 3 shows the distribution of cosine similarity, which indicates the degree of relationship between each of the quizzes, on a page-by-page basis. Three patterns of quizzes that can and cannot be estimated are shown here: the distribution of quizzes that could be estimated (Fig. 2 top), the distribution with some quizzes that could be estimated (Fig. 2 middle), and the distribution of quizzes that could not be estimated (Fig. 2 bottom). The blue vertical dotted line in the graph is the page estimated by the GPT method, and the red vertical dotted line is the answer page. First, as a common pattern among the quizzes that we were able to estimate, the top graph in Fig. 2 shows the highest peak at the point where the quiz was created, and the peaks appear frequently thereafter, though not as frequently as at the point where the quiz was created. In this example, the question was “Who is the new coach?”

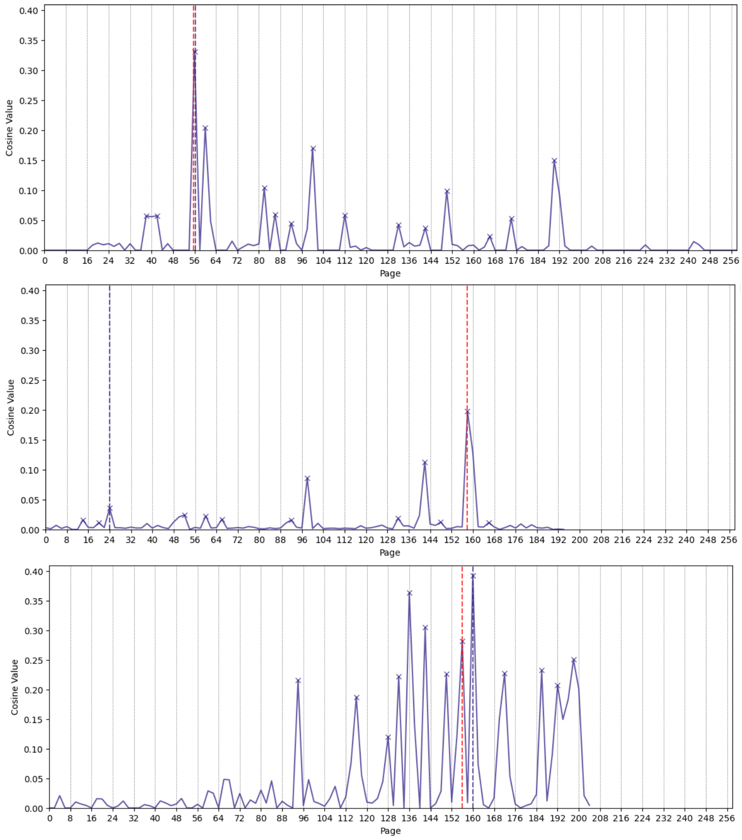
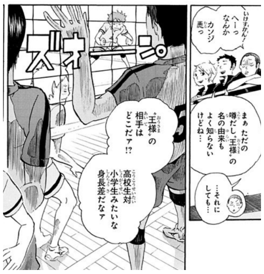
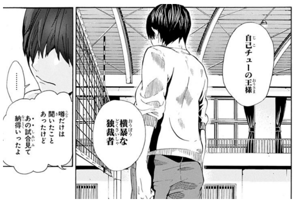
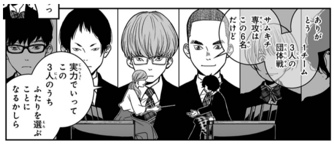
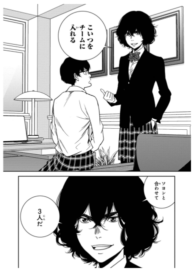


Fig. 2. Cosine similarity distribution and selected pages
(blue vertical dotted line is selected page, red vertical dotted line is answer page).

and the answer was the name of the character. Therefore, the peaks appeared when the character appeared, with the highest peaks appearing on the pages that introduced the character at the time of her first appearance and on the pages where the character was the main topic. The graph in the middle of Fig. 2 shows a pattern in which the graph peaks locally on pages where certain keywords appear. In this graph, the quiz “What does Kageyama Tobio’s nickname, ‘King on the Court,’ mean? (Haikyu!!)” has a peak at the scene where the name of the character mentioned in the quiz and his nickname appear. In this example, the GPT method estimated the highest non-peak scene as the answer page, but other quizzes with similar graph patterns sometimes estimated it correctly. The estimation process of the GPT method will be discussed later. The graph at the bottom of Fig. 2 is an example of a pattern that was common in non-estimated quizzes with an average high value of cosine similarity and a large number of candidate pages. This graph is an example where the answer page is one of the peaks but not the highest in cosine similarity. The question in this example was “Who scored the goal at the end of the game in a 7-on-7 mini-game in the first round of Tokyo City Esperion’s Youth Entrance Test? (Aoashi),” and many pages of the scene where the football game was being played appeared as peaks, with many candidates before and after the answer page. Therefore, estimating quizzes for which the answer is at the beginning peak (top of Fig. 2) is easy, but estimating quizzes for which the answer is at the end of the peak or for which there is a peak before and after the answer (bottom of Fig. 2) may be more difficult. We analyzed the selected pages in quizzes that were not selected correctly and estimated that 74.4% of the quizzes selected the scene before the actual answer page as the answer. This suggests the difficulty of having the middle or end of the page as the answer among the answer candidates. In these non-estimable quizzes, we show an example of a GPT method that includes an output of reasons for estimation.

In the example in Table 5 (No. 1), “What does Kageyama Tobio’s nickname, ‘King on the Court,’ mean? (Haikyu!!)” which is the same quiz introduced in the middle of Fig. 2, the conversation about the rumor of the nickname in the

Table 5. Examples of output results and specific situations when incorrect estimation is performed using GPT method.

	Output results	Scene from estimated page.	Scene from answer page.
No.1	The estimated page number is “24.” The reason for the selection is that a line that appears on page 24 includes the origin of the name “king” and the part where it is expressed as “The king is bad, in addition to violent tossing and turning, blocking and serving, where is the tyrant’s opponent?” This expression is consistent with the answer to the quiz, “Because he is self-centered and a tyrannical dictator,” which is why Kageyama is called “king on the court,” so we determined that this page is the source of the quiz.	 いけすかん! へーっ なんか カンジ 悪っ まぁ ただの 噂だし“王様”の 名の由来も よく知らない けどね… …それに しても… ズオーン。 “王様”の 相手は どこだァ!? 高校生対 小学生みたいな 身長差だなァ Scene from earlier talking about Kageyama Tobio rumor. © Haruichi Furudate, Haikyu!!.	 自己チューの王様 横暴な 独裁者 …… 噂だけは 聞いたこと あったけど ・・あの試合見て 納得いったよ Scene where origin of rumor is revealed. © Haruichi Furudate, Haikyu!!.
No.2	The estimated page is “134”. The reason for the selection is as follows. The text on this page includes the keyword Samkichi, “These are the six Samkichi majors, but...” and the expression “of the three” can also be found on this page. This is in line with the quiz question, “Who is going to participate in the Samkichi Koshien?”, which is strongly related to the quiz question, “ Who got to participate in Samukichi Koshien?” The names “Azumino Ren,” “Kanzaki Shinpaku,” and “Seoyoung” also appear for the respective characters, which leads us to presume that this page is closely related to the quiz answer.	 ありがとう 1チーム 3人の 団体戦 サムキチ 専攻は この6名 だけど 実力でいって この 3人のうち ふたりを選ぶ ことに なるかしら Discussing players for tournament. © Takahiro Ozawa, Asako Seo, Tokyo Toy Boxies.	 こいつを チームに 入れる ソヨンと 合わせて 3人だ Scene in which players who will participate in tournament are determined. © Takahiro Ozawa, Asako Seo, Tokyo Toy Boxies.

early part of the story was estimated. However, the correct answer is “a self-centered king, a tyrannical dictator,” and the scene is an explanation of the origin of the rumor of ‘King on the Court’. This may be due to the fact that the words “Kageyama Tobio” and “king on the court” are common elements on the estimation page and the answer page, and the difference between the two was not understood in context. In Table 5 (No. 2), a scene in which the characters were discussing who to choose was estimated for the quiz “Who was selected to compete in the ‘Samkichi Koshien’?”. However, the answer page was the final scene where the protagonist called three players to be confirmed. Since they are topics that are mentioned at key points in the story, and quizzes that correspond to situations where content that appeared early in the story is explained or revealed in later pages, it is important to understand the context of these quizzes as a story. However, the method used in this study was word-based candidate selection and estimation based on context from those candidates. This remains a challenge for quiz page estimation when time-series processing and relationships with other pages have to be taken into account.

In the estimation of this study, we had the constraint that it was the first volume of a sports comic. Therefore, some of the quizzes included questions such as “Who was appointed as the new director?” or the result of a game score. These are quizzes about characters that tend to be created in the first volume of comics, and quizzes about the results of games, which are unique to sports comics. Thus, quizzes created in other genres are likely to vary in type and difficulty. Furthermore, for cases in which the quizzes are created for volumes later than the first volume, it is necessary to consider the definition of the characters and the flow of the story, and it is conceivable that the present restriction would not be valid. Nevertheless, we expect that the estimation can be performed in a generic way as long as it contains enough keywords necessary and sufficient for the content related to the story.

As an overall discussion, there are still challenges in understanding the overall flow of the story and selecting the correct page. In addition, as shown in Table 4, the accuracy of the estimation is greatly reduced without the text of the answers. Nevertheless, if answer sentences are included, we consider it possible to estimate content that is as close as possible to the topics mentioned in a quiz. In the next step, we plan to investigate whether estimation can be performed from vague sentences other than those in quiz format and from query-type keywords such as those used in searches, and we will propose methods that will lead to comic memory support.

6. Conclusion

In this study, we utilized a quiz-based comic reflection service to extend the quiz dataset and estimated scenes related to quizzes from the quiz questions and answers. To do this, we extracted multiple elements from the comic. Then, we proposed two estimation methods and investigated the accuracy of the "CS method" and the "GPT method". The results showed that the estimation accuracy of the CS method was 55.1% and that of the GPT method was 66.7%. It was also estimated that scenes containing specific keywords or events were easy to estimate as quiz features, while answers that require an understanding of the chronological sequence or the context of the entire comic were difficult to estimate. In addition, the estimation accuracy varied greatly depending on the presence or absence of the answer text, suggesting that if important keywords such as the answer text were included, the content could have been estimated as close as possible to the topic described in the quiz.

In the future, we will examine whether estimation is possible from vague sentences other than quiz-style sentences and query-type keywords such as search. On the basis of this method, we plan to propose other methods that support searches for reflection on comics.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number JP22K12338.

References

[1] Murakami H., Nagaoka Y., Kyogoku R.. (2018) "Creating Character Networks from Comics Using Frames and Words in Balloons." 7th International Congress on Advanced Applied Informatics (IIAI-AAI).

[2] Lee J., and Yeung C. Y.. (2012) "Extracting Networks of People and Places from Literary Texts." In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation: 209–218.

[3] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T. and Aizawa, K.. (2017) "Sketch-based manga retrieval using manga109 dataset." Multimedia Tools and Applications 76 (20): 21811-218388.

[4] Nguyen Nhu, V., Rigaud, C. and Burie, J.. (2019) "Classifying Personalities of Comic Characters Based on Egograms." 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) 1: 44-49.

[5] Chu, W. T. and Li, W. W.. (2019) "Manga face detection based on deep neural networks fusing global and local information." Pattern Recognition 86: 62-72.

[6] Tolle, H. and Arai, K.. (2011) "Method for Real Time Text Extraction of Digital Manga Comic." International Journal of Image Processing 4: 669-676.

[7] Chen, J., Iwasaki, R., Mori, N., Okada, M. and Ueno, M.. (2019) "Understanding Multilingual Four-Scene Comics with Deep Learning Methods." 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) :32-37.

[8] Park, B., Ibayashi, K. and Matsushita, M.. (2018) "Classifying Personalities of Comic Characters Based on Egograms." The 4th International Symposium on Affective Science and Engineering: B3-2.

[9] Nguyen, NV., Rigaud, C., and Burie, JC.. (2019) "Comic MTL: optimized multi-task learning for comic book image analysis." International Journal on Document Analysis and Recognition (IJDAR): 265–284.

[10] Vivoli, E., Baeza, L.J., Valveny Llobet, E., and Karatzas, D.. (2024) "Multimodal Transformer for Comics Text-Cloze." arXiv: 2403.03719.

[11] Guo, H., Wang B., Bai, J., Liu, J., Yang, J., and Li, Z.. (2023). "M2C: Towards Automatic Multimodal Manga Complement." Findings of the Association for Computational Linguistics: EMNLP 2023, 9876–9882.

[12] Zachariah, A., and Rao P.. (2023). "Video Retrieval for Everyday Scenes With Common Objects." ICMR '23 : 565–570.

[13] Baraldi, L., Grana, C., and Cucchiara, R.. "Scene-driven Retrieval in Edited Videos using Aesthetic and Semantic Deep Features." ICMR '16 : 23–29.

[14] Qi, M., Wang, Y., and Li, A.. "Online Cross-Modal Scene Retrieval by Binary Representation and Semantic Graph." MM '17 : 744–752.

[15] kha-white, (2024) “Mokuro. ” URL : https://github.com/kha-white/mokuro.

[16] Li, J., Li, D., Xiong C., Hoi, S.. (2022) “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.” 162(39): 12888-12900.

[17] Roboflow. (2024) “comic-panel-detectors API” URL : https://universe.roboflow.com/personal-ov9jg/comic-panel-detectors/model/7.

[18] OpenAI. (2024) “GPT-4 TurboAPI documentation.” OpenAI. Retrieved from https://www.openai.com/.