A Method to Annotate Who Speaks a Text Line in Manga and Speaker-Line Dataset for Manga109

Tsubasa Sakurai[1], Risa Ito, Kazuki Abe and Satoshi Nakamura

School of Interdisciplinary Mathematical Sciences, Meiji University, Japan
[1] ev190522@meiji.ac.jp

Abstract. Speaker estimation in a manga is one of the components that needs to be recognized in conducting research using manga. To identify the speaker of a text line in a manga, a dataset of who speaks the lines is needed. In order to construct such a dataset easily, we proposed a method to annotate who speaks a text line based on characteristics of information design and the human factor. Then, we developed a prototype system and constructed a dataset that mapped between text lines and speakers in the Manga109 dataset and distributed the dataset on the Web. In addition, we analyzed the dataset and showed that the perfect match rate was about 80% when there were five annotators. It was also found that variation in annotation occurred even with human judgment and that this was partly due to lines requiring reference to other frames. We also found that it was difficult for annotators to map speakers in scenes involving science fiction and battles by calculating the Evaluation Consistency Indicators.

Keywords: Comics, Lines, Speakers, Dataset Construction.

1 Introduction

According to the Japanese E-book Business Research Report 2020 [1], the market size of the e-book is increasing yearly, and the market share of comics is more than 80% in Japan. In addition, sales of e-comics surpassed sales of print comics in 2017. As of 2021, the market size of e-comics has been expanding due to COVID-19 and other factors, and enjoying manga as e-comics is becoming more common. With the spread of e-comics, there will be more and more ways to use and enjoy digital comics.

Research on processing and systems that take advantage of the fact that e-comics are available on digital terminals is also increasing. Mantra [2] performs automatic contextual translation based on image and text information of comics. Other studies have also taken into account the content of comics, such as comic searches [3], recommendations, and spoiler prevention [4]. For these studies, it is necessary to recognize the various elements of comics accurately (e.g., the area of comic frames, the area of lines, the content of lines, onomatopoeia and mimetic words, name and face of a character, facial expressions, the speaker of the lines, and relationships between characters).

To promote such research and development, datasets annotated by many people are essential, and one of these is the Manga109 dataset [5][6], which contains 109 comics drawn by professional cartoonists, with annotations. Other examples include the *eBDtheque* dataset by Guérin et al. [7], the *COMICS* dataset by Iyyer et al. [8], and the four-frame manga dataset for the understanding story by Ueno [9]. As described above, many datasets on comics are available to the public and are used for various research purposes.

One of the issues in translating comics and performing content-based searches and recommendations with high accuracy is recognizing who speaks a line clearly. To increase the accuracy of the recognition system, it is very important to prepare a large dataset of who speaks the lines (hereafter, we call this "speaker-line dataset"). However, there are not enough speaker-line datasets, and it is not easy to construct such a dataset.

In this study, we propose and develop a system to construct a speaker-line dataset easily by dragging a text line and dropping it into a face area of a character. Then, we use our system for the Manga109 dataset and construct a speaker-line dataset. We also analyze the dataset based on multiple indices to clarify fluctuations in human evaluation and situations in which judging a speaker is difficult in the annotation.

The contributions of this paper are as follows:

- This paper proposes and develops a new annotation system to generate a speaker-line dataset.
- This paper constructs a speaker-line dataset with at least four annotators annotating each comic in the Manga109 dataset and distributes the dataset.
- This paper clarifies the difficulties of generating a speaker-line dataset.

2 Related Work

There are various studies on techniques to recognize the elements that compose comics. Nguyen et al. [10] reexamined the definition of frames in comics and proposed a method for extracting them. Wang et al. [11] proposed to extract frames in comics and achieved high performance, accuracy, and no margins. Dubray et al. [12] automatically detected candidate speech balloons and segmentation using machine learning and achieved an F1 score of more than 0.94. Chu et al. [13] proposed a method for character face recognition, and Tolle et al. [14] proposed a method for line recognition with high accuracy. These studies used comic images and text information for recognition, and it can be said that simpler and more accurate methods are being established. A dataset that will provide the correct understanding of comics is needed to promote such research further. This study contributes to such research.

As a technique to promote the research and development of comics, Chen et al. [15] proposed an algorithm for understanding multilingual four-scene comics. Park et al. [16] analyzed the characteristics of characters to realize a comic retrieval system using the characters in comics. Such research is necessary for developing a system that considers the content of comics, and one of the techniques for understanding manga is judging the speaker. Rigaud et al. [17] proposed a method for speaker estimation based on the distance from the tail of a speech balloon. To establish a methodology, these studies need to analyze the relationship between lines and speakers in comics. This study analyzes the relationship between lines and speakers in comics to clarify comic characteristics for highly accurate speaker estimation.

3 A Method to a Construct Speaker-Line Dataset

Constructing a speaker-line dataset is not easy because there are many speakers and a huge number of text lines in a comic. For example, in the Manga109 dataset, the total number of text lines is 147,387 in 109 comics, the average number of speakers is 31.7, and the maximum number of speakers in a comic is 124. We propose a new method to construct the speaker-line dataset easily and quickly.

In comics, a speech balloon and its text line are often settled near the associated speaker because of the aspect of information design. Rigaud et al. [17] used this aspect to associate a speech balloon and its speaker automatically. Here, Fitts's law [18] and its extension to 2D [19] showed that in the action of a user pointing to a certain location, a user could point to a large target with a short distance more quickly and accurately compared with a smaller target with a long distance. Therefore, we can say that mapping between a text line and a speaker for a certain degree of fast and accuracy is suitable for mouse operation.

Considering these characteristics, we propose a method that enables users to map between a text line and a speaker by dragging a text line area and dropping it onto the face area of the speaker. We also developed our system using JavaScript, PHP, and MySQL.

Figs 1, 2, 3, and 4 show our prototype system. In Fig. 1, the annotator is dragging the text line to the face area of the speaker. When the annotator drops the text line onto this face area, The system registers this text line's speaker is this character to the dataset. Our system also considers that the speaker of the line does not appear on the same page when assigning annotations, and the user can select the speaker from the speaker list if necessary (see Fig. 2). The annotator selects "unknown" from the speaker list if the speaker is unknown (Shown in Fig. 2 by the "?" mark). A line that is not considered to be spoken by the speaker, such as explanations of situations and annotations, is assigned as "narration" (Shown in Fig. 2 by the "microphone" mark). In addition, considering the case where the speaker of the previous page can be identified by reading through the comic, annotation can be assigned while keeping track of the comic through the page-by-page management interface (see Fig. 3). Also, since the dataset construction covers a vast amount of content (109 books), a UI (User Interface) for selecting annotation targets is also provided (see Fig. 4).

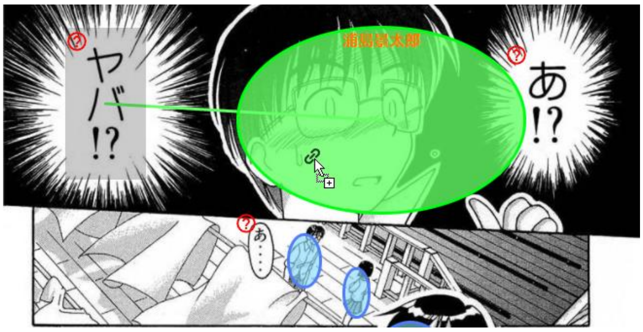


Fig. 1. An annotator can associate a text line with a speaker by dragging a text line area and dropping it onto a face area of speaker. When the annotator starts dragging a text line on the comic, the system shows all the drop targets reference to the face area or body area of speakers. In addition, when the annotator moves the mouse cursor to the drop target, the system changes the color of the drop target and shows the speaker's name.

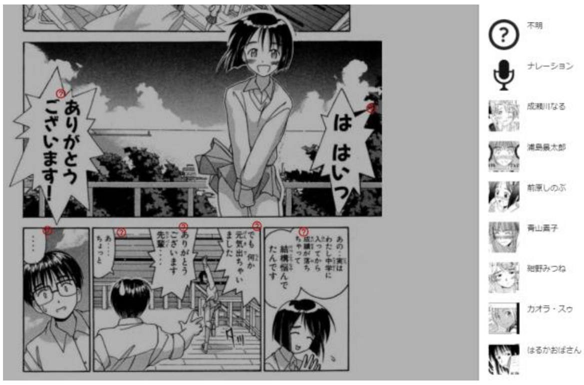


Fig. 2. An annotator can also select a speaker from a speaker list in the comic. This function is useful when the speaker of a text line does not exist on the same page.

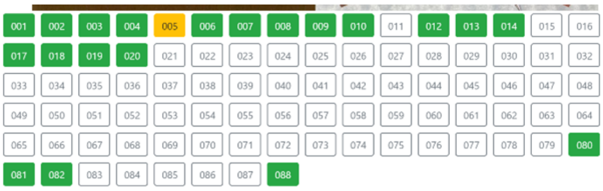
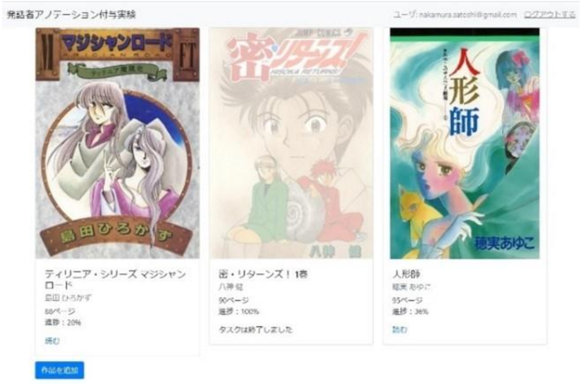


Fig. 3. The system shows the page progress by changing its color.



ティリニア・シリーズ マジシャンロード
島田 ひろかず
88ページ
進捗：20%
読む

密・リターンズ！ 1巻
八神 健
90ページ
進捗：100%
タスクは終了しました

人形師
穂実 あゆこ
95ページ
進捗：36%
読む

Fig. 4. Annotation target selection user interface.

4 Speaker-Line Dataset for Manga 109

We constructed the speaker-line dataset using our system based on the Manga109 dataset. The constructed dataset is available on the webpage[1] of Satoshi Nakamura's laboratory.

Fifty-six annotators contributed to the construction of the dataset. The total number of speaker-line pairs assigned by these 56 annotators was 749,856, and considering that the total number of lines in all comics in Manga109 is 147,918, there were, on average, about five annotators for each line. Furthermore, each comic was annotated by at least four annotators.

Fig. 5 shows the number of annotations by each annotator. In this figure, the horizontal axis indicates the ID of the annotator who performed the annotation (sorted in descending order by the number of annotations). The vertical axis shows the number of annotations by that annotator.

The results show that the two annotators annotated nearly 150,000 lines each. The total number of lines in the target comics is 147,918, indicating that the annotators annotated almost all the lines in the works. These results suggest that our system can adequately map lines to speakers, even with a large number of annotation targets. On the other hand, since more than half of the annotators hardly annotated anything, those annotations may not be helpful.

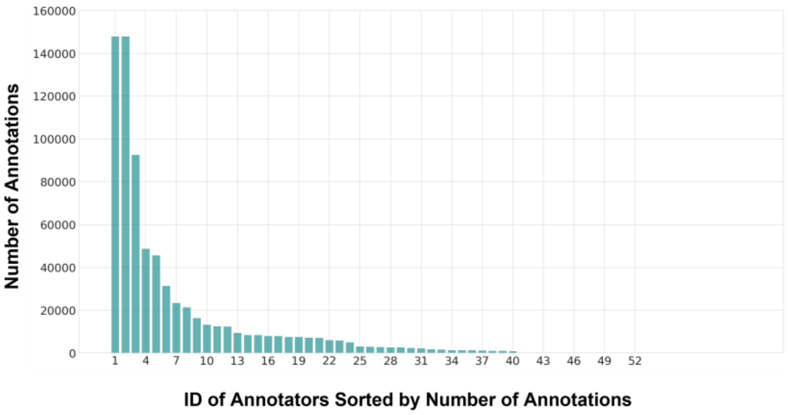


Fig. 5. Number of annotations assigned to each annotator

5 Analysis of the Dataset

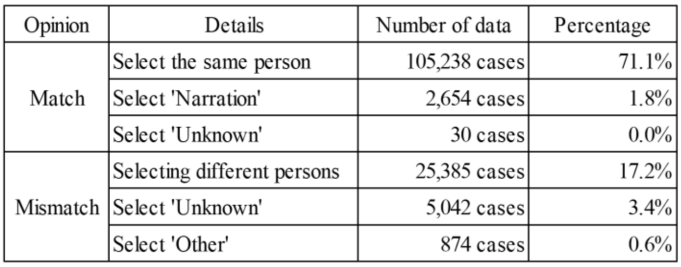
5.1 Agreement Rate of the Annotations and Features of each Frame

Table 1 shows the agreement of annotators' ratings for each line on this dataset. The proportion of all annotators choosing the same character for one line was 71.1%.

Table 2 shows the number and percentage of characters present with the target lines in the same frame. The presence or absence of a candidate speaker in the frame containing the line is also shown. The presence or absence of a candidate speaker in a frame is determined by whether or not the corresponding line and character are depicted within ±50px of each frame. The candidate speaker was the character with the highest number of annotations and was judged on whether it was common to the character that appeared in the frame.

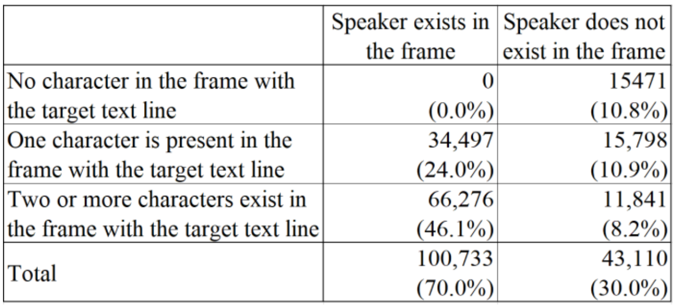
The results showed that the percentage of choice, "Speaker does not exist in the frame," was 30.0%. This result means that 30.0% of text lines do not appear with their speakers in the same frame. Therefore, the recognition system should find a speaker of a target text line for other frames or other pages. In addition, 34,497 cases (68.6%) of "One character is present in the frame" were classified as candidate speakers. On the other hand, 66,276 cases (84.8%) of "Two or more characters exist in the frame" were candidate speakers. This result indicates that if the number of characters per frame is low or zero, there are more opportunities to refer to other pages in the manga.

Table 1. Results of annotation assignment



Opinion	Details	Number of data	Percentage
Match	Select the same person	105,238 cases	71.1%
	Select 'Narration'	2,654 cases	1.8%
	Select 'Unknown'	30 cases	0.0%
Mismatch	Selecting different persons	25,385 cases	17.2%
	Select 'Unknown'	5,042 cases	3.4%
	Select 'Other'	874 cases	0.6%

Table 2. Number of characters and speaker candidates in each frame



	Speaker exists in the frame	Speaker does not exist in the frame
No character in the frame with the target text line	0 (0.0%)	15471 (10.8%)
One character is present in the frame with the target text line	34,497 (24.0%)	15,798 (10.9%)
Two or more characters exist in the frame with the target text line	66,276 (46.1%)	11,841 (8.2%)
Total	100,733 (70.0%)	43,110 (30.0%)

5.2 Relationship between Perfect Match Rate and Number of Annotators

In order to examine how many people should be assigned an annotation, we compare the percentage of the perfect match rate with the number of assignees. Here, the perfect match means that all of the annotators associate a target text line with the same speaker.

Fig. 6 showed the percentage of the perfect match rate when the first N (N=1~6) annotators annotated each of the lines annotated by four or more annotators, five or more annotators, and six or more annotators. The horizontal axis of the figure indicates the situation in which the first N annotators (the first two annotators, the first three annotators, etc.) annotated the lines, and the vertical axis of the figure indicates the perfect match rate. The results show that the graphs are almost the same for all conditions (4, 5, and 6 annotators) and that the perfect match rate gradually decreases as the number of annotators increases.

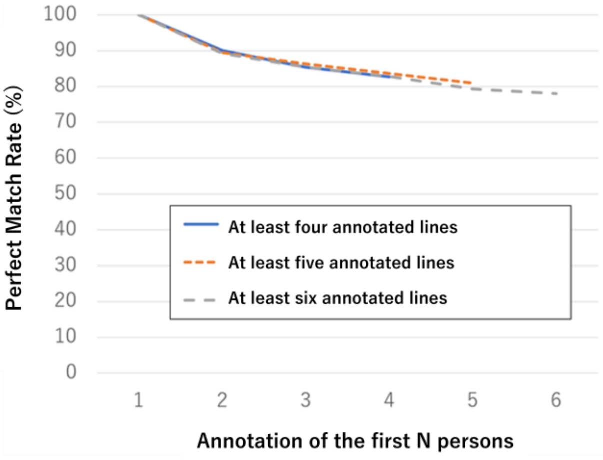


Fig. 6. The perfect match rate of the Nth annotator

5.3 Evaluation Index and Results for Speaker Mapping

In calculating the consistency for the mapping between the lines and the speaker, if everyone's evaluation is a perfect match, then there is no problem. Still, the perfect match is not necessarily considered appropriate as an evaluation index. Even in situations where there is no match, if a certain line is annotated as two characters, it is still good; if it is annotated as three or four characters, the difficulty is considered much higher.

We, therefore, introduce two indices: "Variation," the number of speakers who were mistaken for a given line, and "Max_Match," the maximum number of annotators who were matched for a given line. "Variation" is simply the number of speaker types annotated for a given line, and "Max_Match" is the maximum value of the ratio calculated by dividing the number of annotators for each annotation target by the total number of annotators. For example, if a character A was annotated by eight annotators and character B was annotated by two annotators for a given line, "Variation" is two and "Max_Match" is 0.8. In these items, only the annotations for characters appearing in the comic are included in the classification, and the items classified as narration, other, or unknown are eliminated from the evaluation.

Next, we defined the following ECI (Evaluation Consistency Indicators) for the data in the above dataset (Equation (1)).

$$ECI = \frac{2 \times \text{Max\_Match} \times \frac{1}{\text{Variation}}}{\text{Max\_Match} + \frac{1}{\text{Variation}}} \tag{1}$$

The ECI is the harmonic mean of Max_Match and Variation and is an index that indicates the blurring of evaluations in annotation assignments. In this data, if the value of Max_Match is low and the value of Variation is high, it is considered that the difference in evaluation between people is large, and the value of this indicator becomes low. Therefore, if the value of the ECI is high, it is easy to annotate the line. In contrast, it is difficult to annotate the line if the value is low.

Fig. 7 shows the average of the values of the ECI for each manga, derived from the values of the ECI for all the lines in Manga109. The result shows that there is no manga with a value of 1.0. On the other hand, no manga falls below 0.8.

To confirm whether the ECI work effectively, two works (ARMS, Joouari) were selected, and the average of the values of the ECI on each page was calculated, as shown in Figs. 8 and 9. In these figures, the horizontal axis indicates the number of pages, and the vertical axis indicates the average of the ECI. From these graphs, it can be seen that the value of the ECI varies greatly from page to page.

Figs. 10 and 11 show some of the pages in Figs. 8 and 9, for which the value of the evaluation agreement index was less than 0.7. These results indicate that it is difficult for the annotator to match lines to speakers in battle scenes, dark scenes, and scenes involving spaceships with multiple passengers.

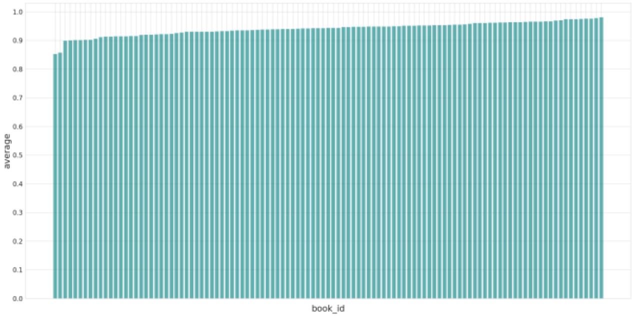


Fig. 7. Distribution of ECI in each comic

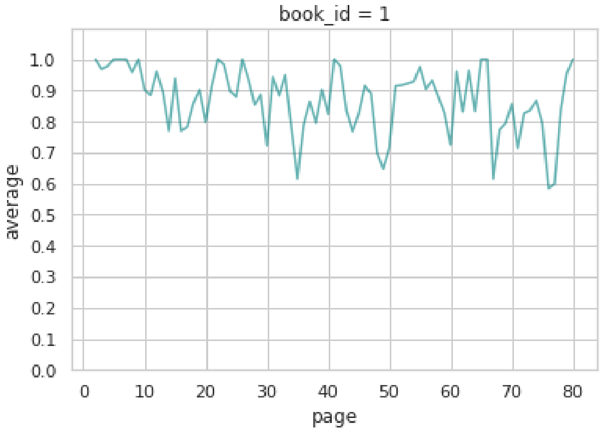


Fig. 8. Distribution of ECI in ARMS

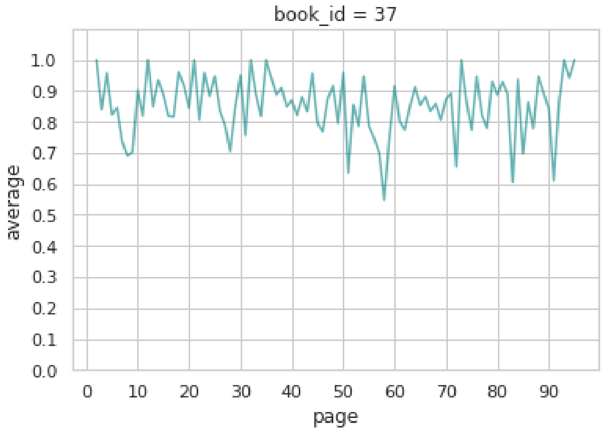


Fig. 9. Distribution of ECI in Joouari7

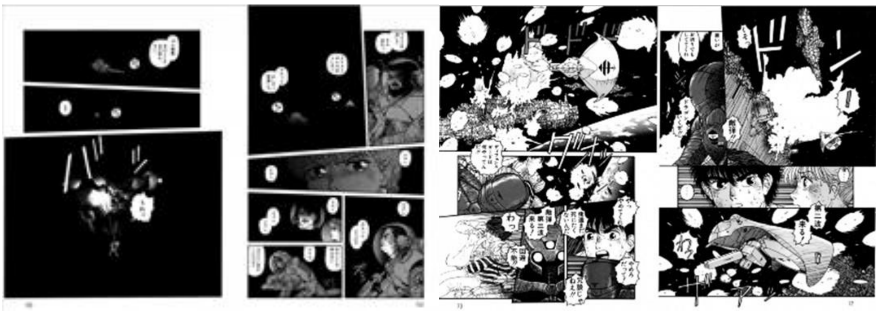


Fig. 10. Battle scenes and dark scenes are difficult to map. © Masaki Kato, ARMS

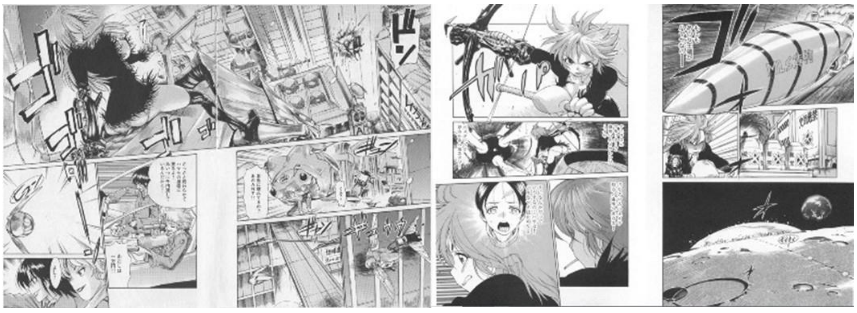


Fig. 11. Battle scenes are difficult to map. © Masakazu Ooi, Joouari

6 Discussion and Prospects

As shown in Tables 1 and 2, the agreement rate for annotations was 71.1%, and the smaller the number of characters in the frame of the target lines, the lower the percentage of candidate speakers present. Therefore, it can be seen that variation in annotation occurs even in human judgment. The reason for the variation in annotation may be due to the fact that there are no speakers in the frame and the lines need to be referred to other frames.

Fig. 6 shows that the perfect match rate decreased by about 10% when comparing the case with two annotators and the case with five annotators. This result indicates that the evaluation is not appropriate just because the speaker's mapping to a line was the perfect match when there were two annotators. Specifically, the perfect match rate differs by 10% between the first two annotators and the first five annotators. Therefore, the appropriate number of annotators needs to be carefully considered.

We defined ECI that were not based on the perfect match rate, and the average of the values of the ECI for each manga is shown in Fig. 7. The results show scenes in which the ECI are low in all manga, as there are no values close to 1.0 or 1.0. The ECI tended to be particularly low in genres such as Science Fiction and Battle, and it was found that there were pages (scenes) with very low values, as shown in Figs. 8 and 9. In addition, it was often difficult to grasp the situation in these specific scenes, such as battle scenes and dark scenes, as shown in Figs. 10 and 11. In these scenes, the following features were observed: the non-existence of characters or the existence of multiple characters in the frame, the existence of many expressions indicating internal speech or physical states in the lines, the non-existence of tails in the speech balloons, the tails not pointing toward the characters, and the non-existence of the speech balloons themselves. In summary, in constructing the speaker-line dataset, it is considered necessary to assign only one or two annotators to simple scenes that can be easily judged by anyone and assign a large number of annotators to difficult scenes such as those in Figs. 10 and 11, and to make decisions by a majority vote or other means.

The automatic estimation of annotation difficulty will be studied and realized based on the dataset constructed in this study. Suppose the automatic estimation of annotation difficulty becomes possible. In that case, it is expected to be possible to allocate resources as required and construct datasets effectively, for example, by microtask such scenes, by a single annotator for such scenes, by majority voting with a large number of annotator for such scenes and by using the judgment of skilled people for such scenes.

7 Conclusion

In this study, we proposed a dataset construction method to enable mapping lines and speakers in Manga109 to contribute to the study of the analysis and understanding of comics. Then, we developed a prototype system and constructed a dataset that mapped between text lines and speakers in the Manga109 dataset and distributed the dataset on the Web. Here, we constructed a huge dataset by 56 annotators to 749,856 annotations. We analyzed the dataset to determine the rate of annotation agreement, the presence or absence of candidate speakers, and how the perfect match rate changed when the number of annotators increased. Furthermore, we set the new ECI not based on the perfect match rate, and analyzed scenes in which it is difficult to estimate the speaker by using these indicators manually. As a result, we found that the ECI was lower in battle scenes and scenes in which it was difficult to grasp the state of the scene like dark scenes in genres such as Science Fiction and Battle. The characteristics of the frames and lines of dialog in these scenes were also clarified.

In the future, we aim to contribute toward the automatic judgment of the speaker by clarifying the difficulty level of annotation for each line. We also plan to clarify differences in human evaluation and to consider methods for dynamically changing the number of annotations required by humans.

8 Acknowledgments

JSPS Grant-in-Aid partly funded this work for Scientific Research JP20K12130.

References

1. Impress Research Institute: E-book Business Research Report 2020 in Japanese, https://research.impress.co.jp/report/list/ebook/500995, last accessed 2021/10/20.
2. Mantra, https://mangahot.jp/, last accessed 2021/10/20.
3. Byeongseon Park, Kahori Okamoto, Ryo Yamashita, Mitsunori Matsushita.: Designing a Comic Exploration System Using a Hierarchical Topic Classification of Reviews. In: Information Engineering Express, Vol.3, No.2, pp.45-57, (2017).
4. Maki, Y, and Nakamura, S.: Do Manga Spoilers Spoil Manga? In: The Sixth Asian Conference on Information Systems, pp.258-262, (2017).
5. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T. and Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. In: Multimedia Tools and Applications, Vol. 76, No. 20, pp.21811-218388, (2017).
6. Ogawa, T., Otsubo, A., Narita, R., Matsui, Y., Yamasaki, T. and Aizawa, K.: Object Detection for Comics using Manga109 Annotations, arXiv:1803.08670, (2018).
7. Guérin, C., Rigaud, C., Mercier, A., AmmarBoudjelal, F., Bertet, K., Bouju, A., Burie, J., Louis, G., Ogier, J. and Revel, A.: eBDtheque: A Representative Database of Comics. In: 12th International Conference on Document Analysis and Recognition, pp. 1145-1149 (2013).
8. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., BoydGraber, J., Daumé III, H. and Davis, L.: The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7186-7195, (2017).
9. Miki Ueno.: Four-Scene Comic Story Dataset for Software on Creative Process, New Trends in Intelligent Software Methodologies, Tools and Techniques, Vol.303, pp.48-56, (2018).
10. Nguyen Nhu, V., Rigaud, C. and Burie, J.: What do We Expect from Comic Panel Extraction? In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Vol. 1, pp. 44-49 (2019).
11. Wang, Y., Zhou, Y. and Tang, Z.: Comic frame extraction via line segments combination. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp.856-860 (2015).
12. Dubray, D. and Laubrock, J.: Deep CNN-Based Speech Balloon Detection and Segmentation for Comic Books. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp1237-1243 (2019).
13. Chu, W. T. and Li, W. W.: Manga face detection based on deep neural networks fusing global and local information, Pattern Recognition, Vol. 86, pp. 62-72.
14. Tolle, H. and Arai, K.: Method for Real Time Text Extraction of Digital Manga Comic, International Journal of Image Processing, (2011).
15. Chen, J., Iwasaki, R., Mori, N., Okada, M. and Ueno, M.: Understanding Multilingual Four-Scene Comics with Deep Learning Methods. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), pp. 32-37(2019).
16. Park, B., Ibayashi, K. and Matsushita, M.: Classifying Personalities of Comic Characters Based on Egograms. In: Proc. the 4th International Symposium on Affective Science and Engineering, and the 29th Modern Artificial Intelligence and Cognitive Science Conference, 2018.
17. Rigaud, C., Le Thanh, N., Burie, J., Ogier, J., Iwata, M., Imazu, E. and Kise, K.: Speech balloon and speaker association for comics and manga understanding. In: 13th International Conference on Document Analysis and Recognition (ICDAR), pp.351-355, (2015).

18. Fitts, P. M.: The information capacity of the human motor system in controlling the amplitude of movement, Journal of Experimental Psychology, pp. 381-391, (1954).
19. MacKenzie, I. S., and Buxton, W.: Extending Fitts' law to two-dimensional tasks. In: Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI '92, pp. 219-226. New York: ACM, (1992).