Question Type Classification for a Comic QA System

Byeongseon Park[†] **Mitsunori Matsushita**[‡]
[†]Graduate School, Kansai University
k281401@kansai-u.ac.jp
[‡]Kansai University
mat@res.kutc.kansai-u.ac.jp

Abstract

In this paper, we perform a classification of the types of questions as a basic examination of a question answering (QA) system for comics. Since comics are multimodal, and hence use texts and illustrations cooperatively, it is difficult to access the information they include. Therefore, we aim to realize a Comic QA system for this purpose. The questions that should be handled by a Comic QA system vary significantly in comparison with conventional QA systems. To solve this problem, we classified the questions into the following two query types: questions requiring bibliographic information (five types) and questions requiring content information (six types). These types are determined by the result of previous work and question sentences collected from Web sites. We performed an automatic classification based on the classification. As a result, we observed that the classification accuracy was high in questions requiring bibliographic information, but low for questions requiring content information.

Keywords: Comic Computing, QA system, Random Forest, Machine learning, Classification

1 Introduction

In recent years, electronic devices such as smart phones and tablets have been rapidly spreading. Along with that, digitization of comics is progressing rapidly.

Comic digitization freed comics from physical constraints as a paper medium [5]. As a result, it is expected that not only a new expression of comics (e.g., switching the contents according to the expansion of the story, giving motion to a frame) [2; 4] will be possible but also flexible information access to comic content (e.g., switching the language according to the native language of the reader, searching for the page where a specific character appears) [7; 11] will be possible.

In this research, our aim is to realize a question answering (QA) system for comics on the premise that flexible information access is possible. Currently, the QA system is one of the research themes being vigorously advanced in the field of natural language processing and we aim to extend it to the QA system for comic content.

Comics are multimodal and composed of multiple elements such as texts and illustrations. For this reason, it is dramatically more difficult than analyzing only text. For example, if the user asks "Which volume did *If box* first appeared in Doraemon?" The QA system that we are aiming for should respond with 'It is *volume 11*." as text. Further, in the case of "I want to see the page where Teacher *Anzai* said that *If you give up, the match is over* in *Slam Dunk*" or "I want to see the scene wherein *Conan* is on the skateboard in *Case Closed*." For such questions it is not enough to answer with text only, presenting a part of the comic page or story also is appropriate. Therefore, to implement such a system, it is necessary to switch the format of the answer to be generated and identify the appropriate part of the comic contents according to the various questions posed by the user.

In this paper, we have focused on questions based on our understanding of the technology that forms the basis of QA to design and implement a Comic QA system. We collected questions about comics and attempted to classify the question types.

2 Related research

2.1 QA Technology

Unlike a general search engine, QA system refers to a technique of presenting a direct answer to a question. The history of QA research is old, and research [3] on technology to retrieve data satisfying the condition from a structured database through questions expressed in natural sentences has been done since the 1960s. In recent years, QA has been extensively studied in the field of natural language processing, given that it is also a technique to retrieve / extract information matching a question from the text data set obtained from the Web and presenting it back to a user. In the following section, the latter question response is handled as a Web question response. The goal of the Comic QA research and implementation is to extend the framework of the aforementioned Web question response. Unlike the web QA system, the Comic QA system does not necessarily apply to this process because it targets multimodal content as described above. Therefore, in the next section, we will organize the process related to the QA system for comics.

2.2 Question Answering for Comics

As mentioned in Chapter 1, in a QA system for comics, there are appropriate scenes that present part of a comic frame or a story rather than text information. In order for the system to recognize a part of a frame or a story as a reply, it is necessary to structure the information contained in the comic (e.g., character, script) to render it machine calculable. By making it possible to calculate the content of comics, if there is a request such as, "I want to see the scene where Conan is on the skateboard in *Case Closed*," the system will receive information such as "*Conan*" and "Skateboard" and it is possible to identify and present the frame containing the keywords extracted from the question. Approaches to structuring the information in comic bibliographies are being carried out by Nomura *et al.* [6], Morozumi *et al.* [9], and it is said that it is possible by using Wikipedia[1] and DBpedia[2].

On the other hand, information regarding comic content is insufficient if taken only from the sources mentioned above, hence it is necessary to extract and structure the information of each comic in more detail. Mito *et al.* are attempting to construct a foundation for answering various questions by manually structuring comic content information [8]. However, it is extremely costly to manually organize the information of all the frames, therefore automatic extraction of comic information is required. An automatic content extraction method for comics would be expected to use image recognition. Sun *et al.* is studying a method of recognizing and identifying the characters of comics using image recognition technology [11]. Although it cannot be said that the correct rate of answer identification at present is necessarily high, by developing this technology, it is expected that specific information (e.g., characters, tools) in a comic could be automatically identified and extracted, and assigned to a frame.

The Comic QA system proposed in this paper assumes that these technologies are available. Currently, we are preparing structured content information manually in line with Mito *et al.*'s research and are using it to conduct our research [8]. Although research on the type classification of general QA has been conducted (e.g. [12]), comics contain multimodal content in which text information and image information are used complementarily as described in Chapter 1. Therefore, there is a possibility that a question type that does not correspond to the general question type classification may appear.

In a QA system, it is necessary to first have a means for determining what the question is about in relation to the search query entered by the user. Therefore, in the next section, we will describe in detail the efforts concerning the type classification of questions for comics.

2.3 Question Type Classification for Comic

Currently, information on comics can be obtained from e-book sales sites (e.g., ComicCmoa[3]) and the like. In such sites, it is generally possible to search by bibliographic information such as the title of the comic, author name, or publisher name. However, the current service is insufficient if you want to search for a certain comic according to its content such as searching for a specific scene as a clue [5].

Under such circumstances, a user could obtain a solution to queries regarding the content of a comic from question sites on the Internet such as "Yahoo!, Chiebukuro[4]" or "Oshiete! Goo[5]." However, in many of these sites, it takes time to get answers from other users or no answer is obtained. To solve such a problem, Fukuda *et al.* have adopted the QA framework described in Section 2.1 and are considering the type classi-

[1] http://www.wikipedia.org

[2] http://www.dbpedia.org

[3] http://www.cmoa.jp

[4] http://chiebukuro.yahoo.co.jp

[5] http://oshiete.goo.ne.jp

fication of questions submitted by users to apply to the comic contents [10]. Fukuda el al. gathered 30 questions about comics from each of the following sites: “Yahoo!, Chiebukuro” and “Oshiete! Goo,” and manually performed the question type classification focusing on interrogative words and keywords appearing in the sentences. The question types are listed below:

- Questions about location

  Questions on the number of stories and the number of volume in which specific scenes and stories are recorded

- Questions about characters

  Questions about comic settings, such as the appearance and affiliation of the characters

- Questions about the story or plot

  Questions about specific contents of the entire series or specific stories

- Questions about title

  Questions asking about the title of a comic

- Questions about others

  Questions not applicable to the above types

Although the above classification is a question type classification corresponding to a comic, since it is classified manually, it is necessary to evaluate whether this classification is valid. Therefore, in the next chapter, we will try to classify comic question types automatically using machine learning.

3 Implementation

3.1 Collection of learning data

To determine question type trends, Fukuda *et al.* collected a total of 60 question sentences from “Yahoo!, Chiebukuro” and “Oshiete! Goo,” and classified the resulting question types. However, the *Questions about location* were 52 (86.7%) and accounted for a large number of the questions. Therefore, in this paper, a total of 318 questions[6] were collected from “Yahoo!, Chiebukuro” and “Oshiete! Goo” to obtain better sampling. In addition, the questionnaire collected in this occasion was targeted to factoid type questions whose correct answers could be determined without ambiguity. However, non-factoid type questions requiring subjective answers and for which the correct answer could not be uniquely determined were excluded from the collection. In addition, sentences and words not related to the content of the question but included in the question sentence (e.g., “Hello”, “Answer please”) were removed in advance while taking into consideration their influence on the classification result.

3.2 Verification using Automatic Classifier

In this paper, we use Random Forest [1] for automatic classification of question types. Also, classifiers were created and evaluated using scikit learn[7] from the machine learning library. The Random Forest Method creates multiple decision trees using randomly extracted learning subsample data and explanatory variables based on given learning data, then it finally creates multiple decision trees for each decision tree. It is a method of ensemble learning to estimate which class data belongs to. Implementations are provided in various fields in terms of simplicity of structure, discrimination ability, processing performance for large-scale data and the like. Since the importance of each explanatory variable is based on the decreasing rate of estimation accuracy when an arbitrary explanatory variable is removed and is calculated at the same time, the strength of the relationship between the objective variable and the explanatory variable can also be taken into account. In terms of the parameters used at learning, the default setting provided in the library was adopted. In this experiment, the Bag of Words (BoW) was composed of the morpheme obtained by morphological analysis of the collected question sentences by the morphological analyzer MeCab[8], and the generated vector was taken as the feature of Random Forest.

We conducted 5-fold cross validation using the questions collected in Section 3.1 to evaluate the question type classification of Fukuda *et al.* Table 1 shows the precision, recall, and F-score of the classification results. From Table 1, it was confirmed that the F-score for the *Questions about Location* and the *Questions about the Title* of a comic was high. This is presumed to be because the learning data contained many charac-

[6] questions accessible as of June 1, 2017 were targeted

[7] http://www.scikit-learn.org/stable/

[8] http://www.mecab.sourceforge.net

Table 1. Classification accuracy of question type based on previous research

Question Type	Number	Precision	Recall	F-score
Location	20	0.64	0.45	0.53
Character	16	0.06	0.06	0.06
Story	56	0.29	0.21	.25
Title	203	0.78	0.85	0.81
Other	23	0.32	0.35	0.33

teristic expressions (“Which volume?”, “Please tell me the title”) for the two question types. On the other hand, the F-score for questions about characters and questions about stories was low. This seems to be caused by the different description formats among questions.

The number of questions categorized as “Other” this time was 23 out of 318 total question sentences, accounting for a high percentage. The question categories classified as “Other” included multiple questions such as questions asking about the release date of the comic and questions about the content of comics which did not belong to the question type classification devised by Fukuda et al(e.g., “What does this mean in this comic?”). Based on these results, it has been determined that the question type classification devised by Fukuda *et al.* is insufficient to classify the questions for a Comic QA. Therefore, in the next section, we will examine the classification standards adapted to more diverse question types based on previous research and the questions included in our collection.

3.3 Reexamination of the question type classification

Fukuda *et al.* compared the elements included in the comic with the elements included in the collected question sentences and roughly divided the type of questions into two elements: bibliographic information (e.g., number of volume，title of comic) and content information (e.g., character，story, script). This research also follows Fukuda *et al.* when classifying the comic question types by first classifying them into questions inquiring about bibliographic information and questions requesting information about content. In this paper, the classification was carried out again taking into consideration those questions not falling under the classification of Fukuda *et al.*, and finally we decided to include 11 question types in total, including the bibliographic information type (5 types) and the content information type (6 types). The question type classification is listed below:

- Bibliographic Information Type
  - Questions about the location
  - Questions about the title
  - Questions about the release date
  - Questions about the publication magazine
  - Questions about the author
- Content Information Type
  - Questions about the progress of the story
  - Questions about the interpretation of the story
  - Questions about the theme of the story
  - Questions about thecharacters
  - Questions about the names of objects, tools and skills
  - Questions about the dialogue

In the next chapter, the automatic classification of question sentences using the 11 question types mentioned above is performed, and the classification accuracy is evaluated.

4 Evaluation and Consideration

In this paper, we performed the question type classification in two stages. As the first step, we evaluated the accuracy of the classification into two types, the bibliographic information type and the content information type. The classification results are shown in Table 2. Upon reviewing the results, it was confirmed that bibliographic information type queries were higher in both accuracy and recall rate.

Next, the classification was carried out according to the 11 categories, a total of 5 categories for the bibliographic information type questions and 6 categories for the content information type questions. However, since the 5-fold cross validation was used to perform this analysis, there is a possibility that the result may be affected if mixed question types with less than five question numbers are mixed. Therefore, in this paper, we evaluated nine kinds of question types excluding questions about the publication magazine with less than 5 questions and questions about the author. The results are shown in

Table 3 and Table 4. According to Table 3, the accuracy of the title question type was relatively excellent. As shown in the results of Table 1, this is considered attributable to the fact that a specific sentence expression such as “what is the title of this comic” was contained in many question statements.

Next, classification was carried out by 11 categories, total of 5 categories of bibliographic information type questions and 6 categories of content information type questions. However, since 5-fold cross validation was used to evaluate this analysis, there is a possibility that the result may be affected if mixed question types with less than 5 question numbers are mixed. Therefore, in this paper, we evaluated with 9 kinds of question types excluding question about publication magazine with question number less than 5 and question about author. The results are shown in Table 3 and Table 4. According to Table 3, the accuracy of the question type regarding the title was relatively excellent. As in the result of Table 1, this is considered to be attributable to the fact that a specific sentence expression such as “what is the title of this comic” was contained in many question statements.

On the other hand, in Table 4, *Questions about the theme of the story*, *Questions about the name of the object, the tool, the skill* and *Questions about the dialogue* were 0. The reason it was not possible to classify these types successfully is that there are few expressions characteristic of each type and additionally the number of questions is small. With regard to these question types, it is necessary to increase the number of relevant samples in the future and to re-evaluate.

Also, compared with the bibliographic information type questions, the accuracy of the content information type questions was poor. As a result, it is conceivable that the percentage of colloquial expressions was high in the content information type questions and the proportion of the named entity expressions was high. In the future, in order to improve accuracy, we will adopt a method to convert colloquial expressions into literal expressions and include a method to distinguish named entities.

Furthermore, through analysis using the Random Forest method Table 5 shows the importance given to each word. In the top 10 words, there are many words related to questions of type that could be classified with high accuracy, such as “title”, “volume”, “hero”. However, it was also confirmed that some words such as “of” and “n” were originally included in the classification. In the future, in order to render the data suitable for analysis, it is necessary to perform the process of removing words whose usage is unclear beforehand.

Table 2. Classification accuracy when categorized into bibliographic information and content information

Question Type	Number	Precision	Recall	F-score
Bibliographic Infomation	235	0.82	0.93	0.87
Content Infomation	83	0.67	0.41	0.51

Table 3. Classification accuracy on bibliographic information type questions

Question Type	Number	Precision	Recall	F-score
Location	20	0.42	0.50	0.45
Title	203	0.86	0.96	0.91
Release date	8	0.50	0.38	0.43
Publication magazine	3	–	–	–
Author	1	–	–	–

Table 4. Classification accuracy on content information type questions

Question Type	Number	Precision	Recall	F-score
progress of the story	13	0.47	0.54	0.50
interpretation of the story	35	0.60	0.26	0.36
cause of the story	8	0.00	0.00	0.00
character	16	0.14	0.12	0.13
names of objects	5	0.00	0.00	0.00
dialogue	6	0.00	0.00	0.00

Table 5. Importance of each word in analysis result using random forest method

Top 10 words	Lower 10 words
タイトル (title)	生 (life)
漫画 (comic)	男女 (gender)
巻 (volume)	転校 (transfer)
の (of)	以下 (less than)
ん (n)	試し (test)
何 (what)	ギャグ (joke)
下 (down)	時期 (season)
主人公 (hero)	生活 (life)
こと (thing)	出版 (publish)
方 (Which)	5

5 Conclusion

In this paper, we conducted a basic study to realize a QA system corresponding to comment questions by gathering questions about comics and attempting to classify the question types. Since we conducted our preliminary assessment based on the classification of previous research and since there were many questions classified as other questions, we performed a new question type classification and performed an evaluation. Questions related to bibliographic information and content information were set up as question types to conduct classification experiments. As a result of the experiment, we succeeded in classifying questions about bibliographic information with high precision, while questions regarding content information displayed a low accuracy, with the exception of a few. The reason for this is that the morpheme analysis cannot be performed normally in the description relating to the content information, therefore many words that interfere with the classification are included. In the future, we aim to improve the classification accuracy through pre-processing such as converting question words with many spoken words into written words. In addition, we will study methods for interpreting the intention of questions to provide a Comic QA system for practical use.

Acknowledgments

This research is supported by JSPS KAKENHI grant numbers JP15K12103.

References

[1] L. Breiman: Random Forests, *Machine Learning*, Vol. 45, pp. 5–32, 2001.

[2] T. Tanaka, K. Shoji, F. Toyama, J. Miyamichi: Layout Analysis of Tree-Structured Scene Frames in Comic Images, In *Proc. 20th International Joint Conference on Artificial Intelligence*, pp. 2885–2890, 2007.

[3] J. Weizanbaum: A Computer Program for The Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, Vol. 9, No.1, pp. 36–45, 1966.

[4] S. Nonaka, T. Swano, and N. Haneda: Development of "GT-Scan", the Technology for Automatic Detection of Frames in Scanned Comic. *FUJIFILM RESERCH & DEVELOPMENT*, No. 57, pp. 46–49, 2012.

[5] M. Matsushita: Potential of Comic Engineering. *Proc. ARG SIG-WI2*, pp. 63–68, 2013(in Japanese).

[6] S. Nomura, A. Morozumi, M. Nagamori, S. Sugimoto: Metadata framework for Manga: A Multi-paradigm Metadata Description Framework for Digital Comics. *Proc. International Conference on Dublin Core and Metadata Applications*, pp. 61–70, 2009.

[7] C. Rigaud, N.T. Le and J.C. Burie, J.M. Ogier, M. Iwata, E. Imazu, K. Kise: Speech Balloon and Speaker Association for Comics and Manga Understanding. *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition*, No. 5, pp. 351–355, 2015.

[8] T. Mito, R. Shirai, K. Hatano, M. Matsushita: Proposal of Data Format for Extracting Relationships within Comic Data. *ARG SIG-WI2*, pp. 71–72, 2013(in Japanese).

[9] A. Morozumi , S. Nomura, M. Nagamori and S. Sugimoto: Metadata Framework for Manga: A Multi-paradigm Metadata Description Framework for Digital Comics, In *Proc. International Conference on Dublin Core and Metadata Applications 2009*, pp. 61–70, 2009.

[10] M. Fukuda, N, Sirozu, M. Matsushita: A Basic Study on Question-Answering System for Comic Contents *Special Interest Group on Language Sense processing Engineering*, SIG-LSE-C003, pp. 57–62, 2012(in Japanese).

[11] W. Sun, J.C. Burie, J.M. Ogier, K. Kise: Specific Comic Character Detection Using Local Feature Matching. *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, No. 5, pp. 275–279, 2013.

[12] U. Hermjakob: Parsing and Question Classification for Question Answering. *Proceedings of the Workshop on Open-domain Question Answering*, Vol. 12, No. 6, pp. 1–6, 2001.