Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6895–6901
Marseille, 11–16 May 2020

Developing Dataset of Japanese Slot Filling Quizzes Designed for Evaluation of Machine Reading Comprehension

Takuto Watarai, Masatoshi Tsuchiya
Toyohashi University of Technology
1–1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan
{watarai, tsuchiya}@is.cs.tut.ac.jp

Abstract

This paper describes our developing dataset of Japanese slot filling quizzes designed for evaluation of machine reading comprehension. The dataset consists of quizzes automatically generated from Aozora Bunko, and each quiz is defined as a 4-tuple: a context passage, a query holding a slot, an answer character and a set of possible answer characters. The query is generated from the original sentence, which appears immediately after the context passage on the target book, by replacing the answer character into the slot. The set of possible answer characters consists of the answer character and the other characters who appear in the context passage. Because the context passage and the query shares the same context, a machine which precisely understand the context may select the correct answer from the set of possible answer characters. The unique point of our approach is that we focus on characters of target books as slots to generate queries from original sentences, because they play important roles in narrative texts and precise understanding their relationship is necessary for reading comprehension. To extract characters from target books, manually created dictionaries of characters are employed because some characters appear as common nouns not as named entities.

Keywords: Machine Reading Comprehension, Slot Filling Quiz, Novel Characters

1. Introduction

Machine reading comprehension is one of crucial goals of the field of natural language processing. To achieve this goal directly, it is necessary to define what reading comprehension is, however it is quite difficult. Therefore, instead of defining reading comprehension directly, evaluation based on benchmark tasks, which can be achieved if and only if target sentences are precisely understood, are widely employed.

There are three typical benchmark tasks widely employed for evaluation of machine reading comprehension. The first one is the slot filling task (Hill et al., 2016; Hermann et al., 2015) which is defined as follows. Suppose $n+1$ sentences s_i^{i+n+1} which appear sequentially in a target text. When a context passage C which consists of n sentences s_i^{i+n} and a query q generated from the sentence s_{i+n+1}, which appears immediately after the passage C, by replacing a word into a slot, and a set of possible answer words A are given, a machine must select the correct answer word a, which appears in the original sentence s_{i+n+1}, from the set A. The second one is the 4 multiple choice question task (Richardson et al., 2013) which is defined as follows: when a passage C, a query q about a fact described in the passage C and a set of possible answer sentences A are given, a machine must select the correct answer a, which meets conditions given by the passage C and the query q, from the set A. The third one is the question answering task (Rajpurkar et al., 2016). This task is defined as follows: when a passage C and a query q about a fact described in the passage C are given, a machine must extract an answer a, which meets the condition given by the query q, from the passage C. These definitions share the same framework: a machine must answer a quiz about the context passage C, and it is crucial to prepare quizzes which a machine can answer if and only if it precisely understands context passages.

This paper focuses on the slot filling task to reduce human works and human biases. When constructing a dataset of the 4 multiple choice question task, a time consuming work to composite a query q and a set of possible answer sentences A is necessary, even if the passage C is imported from another language resource. Although crowd-sourcing is common approach for such time-consuming work in these days, crowd workers may cause hidden biases described in (Tsuchiya, 2018). Construction of a dataset of the question answering task shares the same weak point, because it is also necessary to prepare a query q and an answer a manually.

When focusing on the slot filling task, the procedure to extract what kind of words from the context passage C and the original sentence s_{i+n+1} as a slot and possible answers is still an open problem. For example, Children's Book Test (henceforth denoted as CBT) (Hill et al., 2016), which is a dataset of English slot filling task widely used in (Cui et al., 2017; Kadlec et al., 2016; Trischler et al., 2016), extracts two kinds of words as slots and possible answers: common nouns and named entities detected by Stanford CoreNLP Toolkit (Manning et al., 2014). Because of its naive method, CBT has two weak points. The first weak point is that CBT contains some trivial quizzes because it does not distinguish named entity categories when extracting slots and possible answers. Fig. 1 shows an example trivial quiz of CBT. Because the prefix "Mr." appears immediately before the slot X in the query q, it is easy to estimate that the slot X is a person name. However, the set A of possible answers contains only two person names such as "Peter" and "Toad". It means that this quiz does not work as a 10 multiple choice question but works as a binary choice question. The second weak point described in (Kaushik and Lipton, 2018) is that the last sentence s_{20} of the context passage C is only required to select the correct answer a from the set A for large portion of CBT. These two weak points strongly suggest that extraction procedure

s_1	"I must dress.
s_2	I'm ever so late," he said, hurrying upstairs; and the princess, with a little sigh, went down to the royal drawing-room.
s_3	CHAPTER II.
s_4	Princess Jacqueline Drinks the Moon.
s_5	The King and the Prince: When dinner was over and the ladies had left the room, the king tried to speak seriously to Prince Ricardo.
s_6	This was a thing which he disliked doing very much.
s_7	"There's very little use in preaching," his Majesty used to say, "to a man, or rather a boy, of another generation.
s_8	My taste was for books; I only took to adventures because I was obliged to do it.
s_9	Dick's taste is for adventures; I only wish some accident would make him take to books .
s_{10}	But everyone must get his experience for himself; and when he has got it, he is lucky if it is not too late.
s_{11}	"I don't like to doubt your word, Mr. Toad," said he, "but you'll have to show me before I can believe that."
s_{12}	Old Mr. Toad's eyes twinkled.
s_{13}	Here was a chance to get even with Peter for watching him change his suit.
s_{14}	"If you'll turn your back to me and look straight down the Crooked Little Path for five minutes, I'll disappear," said he.
s_{15}	"More than that, I give you my word of honor that I will not hop three feet from where I am sitting."
s_{16}	"All right," replied Peter promptly, turning his back to Old Mr. Toad.
s_{17}	"I'll look down the Crooked Little Path for five minutes and promise not to peek.".
s_{18}	So Peter sat and gazed straight down the Crooked Little Path.
s_{19}	It was a great temptation to roll his eyes back and peep behind him, but he had given his word that he wouldn't, and he didn't.
s_{20}	When he thought the five minutes were up, he turned around.
q	Old Mr. X was nowhere to be seen.
a	Toad
A	Peter, Toad, back, first, minutes, peek, people, right, sort, word

Figure 1: An example of trivial quizzes of CBT. Because the prefix "Mr." appears immediately before the slot X in the query q, it is easy to estimate that the slot X is a person name. The set A of possible answers contains only two person names such as "Peter" and "Toad", thus, this quiz do not work as a 10 multiple choice question, but works as a binary choice question. And more, the context sentences s_1^2 belong to CHAPTER I and the context sentences s_3^{20} belong to CHAPTER II. Because there is a big contextual gap between CHAPTER I and CHAPTER II, either a machine or a human worker cannot use the context sentences s_1^2 when selecting a correct answer from the set A. These facts suggest that the size of the set of possible answers and the length of the context passage do not reflect the degree of the difficulty of CBT.

of slots and possible answers crucially affects the quality of the dataset.

This paper focuses on characters in narrative texts as slots and possible answers unlike to CBT, because of two reasons. The first reason is significant roles of characters in narrative texts. The precise recognition of their roles and their relationships is crucial when reading a book, thus, quizzes about them are suitable for evaluation of machine reading comprehension. The second reason is to keep the quality of the dataset. As shown in Fig. 1, if named entity categories are not distinguished when extracting slots and possible answers, a trivial quiz may pollute the dataset. Our approach to extract only characters as slots and possible answers can avoid this risk. Unfortunately, extracting of characters is not a trivial task because many characters do not appear as named entity but as common nouns, and because there are many named entities missed by the existing named entity extractor. In order to resolve the above problems, manually created dictionaries of characters are employed to extract characters in this paper.

2. Construction of Dataset

2.1. Target Books of Dataset

This section describes the target books of our dataset, because selection of the target books is one of the important factors which affect the quality of the dataset.

This paper focuses the books which are written for kids and are published by Aozora Bunko[1] as the language resource of the context passages and the original sentences. There are two reasons of this decision. Aozora Bunko is a group to collect Japanese books whose copyrights have expired and to publish them on the Internet, and its archive is widely referred as a public domain Japanese corpus. The first reason is that its availability is suitable for the language resource of the dataset. For constructing a machine reading comprehension testset, clear text structure and consistent interpretation is important, as already indicated by CBT. The second reason is that books written for kids may have clearer text structures than books written for adults.

To select the books written for kids from the whole books of Aozora Bunko, its metadata is employed. To each book of Aozora Bunko, the 5-tuple metadata is assigned: a ti-

[1] https://www.aozora.gr.jp/

# of total books	15,076
# of novels written for kids	1,093
# of novels written for kids in the current kana orthography	810
# of chapters in the target books	5,285
# of sentences in the target books	246,369

Table 1: Statistics of Aozora Bunko and the target books. This statistics was generated from the Git repository of Aozora Bunko at December 1, 2018.

Author	# of books
Ogawa Mimei (小川 未明)	388
Miyazawa Kenji (宮沢 賢治)	64
Yumeno Kyusaku (夢野 久作)	50
Niimi Namkichi (新美 南吉)	43
Kusuyama Masao (楠山 正雄)	42
Unno Juuzou (海野 十三)	37
Edogawa Rampo (江戸川 乱歩)	33
Toyoshima Yoshio (豊島 与志雄)	30
Takeshisa Yumeji (竹久 夢二)	18
Kosakai Fuboku (小酒井 不木)	14
Total	810

Table 2: Top-10 authors of the target books

tle, an author, a publication year, a classification code and an orthography style. The classification code system of Aozora Bunko is designed based on the Nippon Decimal Classification (henceforth denoted as NDC) system which is widely used in Japanese libraries to represent subjects of books, and the small difference between the system of Aozora Bunko and NDC is the prefix "K" to represent whether a target book is written for kids or not. It means that the code "K913" is assigned to the novels written for kids.
The orthography style metadata, which is assigned to each book of Aozora Bunko, is also important to select the target books. It represents whether a target book is written in the historical kana orthography or in the current kana orthography. Because almost all of existing natural language processing tools for Japanese is designed for texts written in the current kana orthography, it is difficult to handle texts written in the historical kana orthography. We, therefore, eliminate the books written in the historical kana orthography from the target books of our dataset.
According to the above discussion, novels written for kids in the current kana orthography are selected as the target books of our dataset. Table 1 shows the statistics of Aozora Bunko and the target books. The target books are 5.4% (810 books) of the books included in the Git repository [2] of Aozora Bunko at December 1, 2018. The number of authors of the target books is 54, and the top-10 authors of them are shown in Table 2. They are ones of most famous Japanese novel writers who were active from late 19th century to early 20th century.

2.2. Construction Procedure

Our proposing procedure to create a slot filling quiz consists of the following steps.

[2] https://github.com/aozorabunko/aozorabunko.git

> Toshio laughed, and said, "You are off target. Why can we solve this case without interrogating Tomekichi?" Mr. Oda and I were amazed and stared at Toshio's face.
>
> Disappeared evidence
>
> After a while, Mr. Oda asked, "Will you interrogate the idiot?" Toshio laughed like a nasty.

Figure 2: An example of chapter heading. The example is extracted from "Wisdom of Idiot" written by Kozakai Fuboku. A typical chapter heading of Aozora Bunko does not form a complete sentence, but forms a short phrase surrounded by two or more linefeed symbols.

1. Divide a target book into sentences.
2. Extract characters from a sentence.
3. Create a slot filling quiz.

The details of each step are described below.

2.2.1. Formatting of Target Books

This section describes restrictions introduced by chapter structure when selecting a target text from a book.
We introduce the following restriction for our database: a context passage C and an original sentence of a query q must belong to the same chapter. As already described in Section 1., the slot filling task is to select the correct answer for the slot of the query q from the set of the possible answers based on the context information given by the context passage C. Suppose the example quiz shown in Fig. 1. In this quiz, the context sentences s_1^2 belong to CHAPTER I and the context sentences s_3^{20} belong to CHAPTER II. Because there is a big contextual gap between CHAPTER I and CHAPTER II, either a machine or a human worker cannot use the context sentences s_1^2 when selecting the correct answer from the set A. In other words, if the context passage C and the original sentence of the query q contains one or more chapter boundaries, it is impossible to use whole of the context passage C when selecting the correct answer for the slot of the query q. This is the reason of the above restriction.
The above restriction requests that the procedure to extract sentences from target books consists of two steps. The first step is to divide each book to chapters, and the second step is to divide each chapter to sentences. Unfortunately,

s_1	So Mr. Grimes came up to Harthover next day with a very sour face; but when he got there, Sir John was over the hills and far away; and Mr. Grimes had to sit in the outer servants' hall all day, and drink strong ale to wash away his sorrows; and they were washed away long before Sir John came back.
s_2	For good Sir John had slept very badly that night; and he said to his lady, "My dear, the boy must have got over into the grouse – moors, and lost himself; and he lies very heavily on my conscience, poor little lad.
s_3	But I know what I will do."
s_4	So, at five the next morning up he got, and into his bath, and into his shooting-jacket and gaiters, and into the stableyard, like a fine old English gentleman, with a face as red as a rose, and a hand as hard as a table, and a back as broad as a bullock's; and bade them bring his shooting pony, and the keeper to come on his pony, and the huntsman, and the first whip, and the second whip, and the under-keeper with the bloodhound in a leash – a great dog as tall as a calf, of the colour of a gravel-walk, with mahogany ears and nose, and a throat like a church-bell.
s_5	They took him up to the place where Tom had gone into the wood; and there the hound lifted up his mighty voice, and told them all he knew.
s_6	Then he took them to the place where Tom had climbed the wall; and they shoved it down, and all got through.
s_7	And then the wise dog took them over the moor, and over the fells, step by step, very slowly; for the scent was a day old, you know, and very light from the heat and drought.
s_8	But that was why cunning old Sir John started at five in the morning.
s_9	And at last he came to the top of Lewthwaite Crag, and there he bayed, and looked up in their faces, as much as to say, "I tell you he is gone down here!"
s_{10}	They could hardly believe that Tom would have gone so far; and when they looked at that awful cliff, they could never believe that he would have dared to face it.
q	But if the X said so, it must be true.
a	dog
A	boy, keeper, huntsman, dog, under-keeper

Figure 3: An example quiz where characters appear as common nouns. This quiz is created from a part of the book "The Water-Babies" by ourselves. The dog appears as a common noun in the sentence s_4 and the sentence s_7, and it plays an important role as a character in this book. It, therefore, is necessary to understand its role when selecting the dog as an answer from the set of possible answers. The existing named entity extractors cannot extract these characters which appear as common nouns.

data provided by Aozora Bunko contains no explicit chapter boundaries like CHAPTER II shown in the sentence s_3 of Fig. 1. Data provided by Aozora Bunko contains implicit chapter boundaries as shown in the center of Fig. 2. A typical chapter heading of Aozora Bunko does not form a complete sentence, but forms a short phrase surrounded by two or more linefeed symbols. Thus, this superficial clue is employed to detect chapter headings and to divide each book into chapters as the first step. As the second step, each chapter is divided into sentences based on punctuation marks with the exception that punctuation marks which appear in dialogues surrounded by brackets are not used[3]. This exception is introduced to avoid the sentence division error whose example is shown in the sentences s_1^2 of Fig. 1.

2.2.2. Extraction of Characters

This section describes the procedure to extract characters from actual sentences contained in the target books.

Although characters play important roles in narrative texts, extracting characters is not a trivial task. There are two problems when extracting characters from the target books. The first problem is to identify characters among general common nouns. Fig. 3 shows an example quiz where characters appear as common nouns. The dog appears as a common noun in the sentence s_4 and the sentence s_7, and it plays an important role as a character in this book. Therefore, it is necessary to understand its role when selecting the dog as an answer from the set of possible answers. In other words, for evaluation of machine reading comprehension, extracting such common nouns which appear as characters and employing them as slots and possible answers are important. The second problem is detection errors of named entities. Fig. 4 shows examples of named entities missed by the existing Japanese named entity extractor, the combination of JUMAN++ 2.0.0RC2 (Morita et al., 2015) and KNP 4.19 (Kawahara and Kurohashi, 2006). Because almost all of existing Japanese named entity extractors are trained on the newspaper corpus, domain mismatch between their training corpus and the target books may cause many detection errors. These domain-specific named entities appear as characters in the target books, and play important roles. Therefore, it is necessary to extract them and to recognize their roles when reading the target books.

In order to resolve the above two problems when extracting characters from the target books, we manually prepare a dictionary for each book. There are 5 categories prepared for the dictionary entries: person name, location name, organization name, personal entity, and artificial entity. The first three categories, person name, location name and organization name, are prepared to extract named entities missed by the existing named entity extractor. In order to reduce the manual construction cost of dictionaries, named entities extracted by the existing extractor are not registered

[3]In Japanese, the punctuation mark is mainly used as a sentence boundary, and is not used as an end of an abbreviation word.

Type	Examples of not extracted named entities
Person Name	二十面相 (Twenty Faces), 悪魔 (Fiend) , 丁七唱 (Choshichitonou) , アア (Aa), ノロちゃん (Noro), ギネ (Gine), フウフィーヴオ (Fufivo), ペンネンネンネンネン (Pennemnemnemnen), ネネム (Nenemu), エキモス (Ekimosu)
Organization Name	月世界探検隊 (Expedition of Lunar), 少年探偵団 (Boy detectives), イーハトーヴ火山局 (Center of Ihatov volcano), 読売新聞 (Yomiuri Newspaper)
Location Name	S 国 (S country), 六天山塞 (Rokutensansai), 地獄の一丁目 (The first block of hell), ハンムンムンムンムン・ムムネ市 (Hanmunmunmunmun Mumune City), 火星 (Mars)

Figure 4: Examples of named entities which are missed by the existing Japanese named entity extractor, the combination of JUMAN++ 2.0.0RC2 and KNP 4.19. These named entities appear as characters in the target books, and play important roles. Therefore, it is necessary to extract them and to recognize their roles when reading the target books.

to dictionaries, and ones missed by the existing extractor are only registered. The remaining two categories, personal entity and artificial entity, are prepared to register common nouns which appear as characters in the target book.

2.2.3. Construction of Slot Filling Quiz

This section explains the formal construction procedure of a slot filling quiz. We need two constants for the construction procedure: n is the length of the context passage, and m is the size of the set of possible answers.

The construction procedure consists of 4 steps. The first step is to obtain the context passage candidate s_i^{i+n} and the original sentence candidate s_{i+n+1} from the target book. As already described in Section 2.2.1., if either the context passage candidate or the original sentence candidate contains one or more chapter boundaries, the candidate is discarded. The second step is to extract characters from the context passage candidate and to get the set W which consists of all extracted characters. Its size $|W|$ must be equal or larger than m, thus, if W does not meet this condition, the candidate is also discarded. The third step is to extract characters from the original sentence candidate and to get the set U which consists of all extracted characters. Because an answer character must appear both in the context passage candidate and in the original sentence candidate, the candidate is discarded if the intersection $\{U \cap W\}$ is empty. Finally, every member of the intersection is adopted as an answer character a, and a query q is generated from the original sentence candidate by replacing the answer character a into a slot. The set A of possible answer characters is a randomly selected subset of W under the condition that the set A must contain the answer character a.

2.3. Length of Context Passage

This section explains our investigation to determine the length of the context passage C.

In order to evaluate influence of the length of the context passage to the difficulty of the slot filling task, we investigate the ratio of quizzes correctly answered by a human worker. As already described in Section 1., each quiz is defined as a 4-tuple: a context passage C, a query q holding a slot, an answer a, and a set A of possible answers. Table 3 shows the result of our investigation. The row where the length of the context passage is equal to 0 is the accuracy achieved by a human worker when he/she was asked to answer 100 quizzes without context information. The row where the length of the context passage is equal to 20 is the accuracy achieved by a human worker when he/she was asked to answer 100 quizzes with context information given by 20 context sentences. As shown in Table 3, 5 trials where the length of the context passage is 0, 1, 5, 10 and 20 were conducted in our investigation. Table 3 shows that the ratio of the correct answers exceeds 80% when 10 or more context sentences are given. Thus, 10 is adopted as the length of the context passage of our dataset.

Length of Context Passage	Human Accuracy
0	.33
1	.30
5	.58
10	.85
20	.89

Table 3: Results of manual investigation. A human worker achieves 80% when 10 or more context sentences are given. This result suggests that 10 or more context sentences are required to select an appropriate answer from 5 possible answers.

Table 3 also shows that a human worker achieves the accuracy near to 30%, even when either no context sentence or a context sentence is given. We think that these results are caused by trivial quizzes like Fig. 1. Eliminating such trivial quizzes remains as a future task.

3. Construction Result

This section explains the brief summary of the current status of our developing dataset.

Fig. 5 shows an example quiz of our developing dataset. The quiz of our dataset is written in Japanese, and the quiz of Fig. 5 is reconstructed from English version of "The Beats of Mt. Nametoko" written by Miyazawa Kenji. The quiz consists of 4 elements: the context passage C which consists of 10 sentences, the query q which is generated from the original sentence which appears immediately after the context passage C by replacing a character into a slot, the correct answer a, and the set A of possible answers. We adopt 5 as the size of the set of possible answers.

Table 4 shows the statistics of our developing dataset. Out-of-vocabulary words (denoted as OOV) of the validation set and OOVs of the test set are defined against the vocabulary of the training set. Thus, Table 4 shows that the vocabulary of the validation set and the vocabulary of the test set are quite similar to the vocabulary of the training set.

s_1	This is all what I have heard from others, or worked out for myself.
s_2	It may not be entirely true, but I, at least, believe it.
s_3	What is certain, at any rate, is that Mt. Nametoko is famous for its bear's liver.
s_4	It is good for the stomach-ache and it helps wounds to heal.
s_5	At the entrance to the Namari hot springs there is a sign that says Bear's Liver From Mt. Nametoko.
s_6	So it is certain that there are bears on Mt. Nametoko.
s_7	I can almost see them, going across the valleys with their pink tongues lolling out, and the bear cubs wrestling with each other till finally they lose their tempers and box each other's ears.
s_8	It was those same bears that the celebrated bear hunter Kojuro Fuchizawa once killed so freely.
s_9	Kojuro Fuchizawa was a swarthy, well-knit, middle-aged man with a squint.
s_{10}	His body was massive, like a small barrel, and his hands were as big and thick as the handprint of the god Bishamon that they use to cure people's sicknesses at the Kitajima Shrine.
q	In summer, X wore a cape made of bark to keep off the rain, with leggings, and he carried a woodsman's axe and a gun as big and heavy as an old-fashioned blunderbuss.
a	Kojuro
A	Bishamon, Kojuro, Fuchizawa River, Nakayama Highway, Kitajima

Figure 5: Example quiz of our developing dataset. The quiz of our dataset is written in Japanese, and this quiz is reconstructed from English version of "The Beats of Mt. Nametoko" written by Miyazawa Kenji for explanation. The length of the context passage of our dataset is 10, and the size of the set of possible answers of our dataset is 5.

	Training set	Validation set	Test set
# of authors	49	12	10
# of books	390	13	40
# of chapters	1,791	116	203
# of sentences	561,770	44,440	34,188
# of quizzes	19,151	1,547	2,092
# of words	11,958,153	783,926	1,012,154
Size of vocabulary	54,896	11,077	13,537
# of OOVs	—	1,586	1,972
Ratio of OOVs	—	0.143	0.146

Table 4: Statistics of our developing dataset

The statistics of our manually created dictionaries to extract characters are shown in Table 5. As described in Section 2.2.2., there are 5 categories prepared for the dictionary entries: person name, location name, organization name, personal entity, and artificial entity. The first three categories are prepared to extract named entities missed by the existing named entity extractor. The column of the number of extracted characters using the dictionary shows that 1/3 of persons are missed by the existing extractor. The remaining two categories are prepared to identify characters among generic common nouns in the target book. All occurrences of these entries are missed, if the dictionary is not available. Therefore, Table 5 means that the manually created dictionaries are necessary to obtain sufficient coverage when extracting characters.

Table 6 shows the results of the preliminary experiments to evaluate the difficulty of our developing dataset. Two existing NN-based reading comprehension models, which were proposed for CBT in (Cui et al., 2017; Kadlec et al., 2016), are employed for our dataset. Because both models achieved lower performance than the performance of the human worker shown in Table 3, there is enough research space to improve them. The columns of CBT shows the performance achieved by both models for CBT.[4] Attention Sum Reader model achieves comparable performance for both datasets, and Attention on Attention model achieves lower performance for our dataset than CBT. These results suggest that our dataset keeps comparable difficulty to CBT, although the size of the set of possible answers is reduced from 10 to 5.

4. Discussion

This paper described our developing dataset of Japanese slot filling quizzes designed for evaluation of machine reading comprehension. Our dataset has two contributions. The first contribution is that this is the first slot filling dataset focuses on characters of Japanese narrative texts as far as we know. Our dataset consists of automatically generated quizzes from narrative texts which are written for kids of Aozora Bunko, and each quiz is defined as 4-tuple of:

- A context passage C which consists of 10 sentences,
- A query q which is generated from the original sentence appears immediately after the passage C by replacing a character to a slot,
- An answer character a which appears in the original sentence of the query q, and
- A set A which consists of the answer character a and 4

[4]These results for CBT are copied from their proposed papers.

	# of dictionary entries	# of extracted characters using the dictionary	# of NEs extracted by JUMAN and KNP
Total	3,326	79,042	134,080
Person	1,364	50,754	105,019
Location	729	5,739	18,501
Organization	147	1,162	10,560
Personal entity	807	17,923	N/A
Artificial entity	338	3464	N/A

Table 5: Statistics of the manually created dictionaries and extracted characters

	CBT		Our Dataset			
	NE	CN	Total	Location	Organization	Person
Attention Sum Reader (Kadlec et al., 2016)	0.686	0.634	0.652	0.538	0.597	0.668
Attention on Attention (Cui et al., 2017)	0.720	0.694	0.569	0.425	0.459	0.594

Table 6: Preliminary results of existing NN-based models. Two existing NN-models proposed for CBT achieve lower performances than the human worker's performance shown in Table 3.

possible characters who appear in the context passage C and the original sentence of the query q.

The unique point of our dataset is that our dataset focuses on characters of target books as slots to generate queries from original sentences, because they play important roles in narrative texts and precise understanding of their relationship is necessary for reading comprehension. Although characters play important roles of narrative texts, extracting characters is not a trivial task because many characters do not appear as named entities but as common nouns and because there are many named entities missed by the existing named entity extractor. Therefore, a manually created dictionary of characters is employed to extract characters. The second contribution is empirical analysis of the length of the context passage. Our investigation showed that 10 or more context sentences are necessary to select an appropriate character from 5 possible characters. And more, this paper also described the preliminary results of NN-based machine reading comprehension models against our dataset.

Because our dataset is still developing, there are several points must be examined. The first point is to improve its quality. Table. 3 shows that a human can select a right character for 33% of quizzes even if no context passages are given. We think that this is a big fault of the dataset, and plan to fix it as soon as possible. The second one is more detailed analysis of the length of the context passage. Table. 3 also shows that a human cannot achieve 100% accuracy even if 20 context sentences are given. The more detailed analysis of this reason is necessary. In the future, we plan to study a machine reading comprehension model that takes into account omissions using this dataset.

5. Bibliographical References

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. In *Proc. of ACL2017 (Vol. 1: Long Papers)*, pages 593–602.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The goldilocks principle: Reading children 'sbooks with explicit memory representations'. In *Conference paper at ICLR2016*.

Kadlec, R., Schmid, M., Bajgar, O., and Kleindienst, J. (2016). Text understanding with the attention sum reader network. In *Proc. of ACL2016 (Vol. 1: Long Papers)*, pages 908–918.

Kaushik, D. and Lipton, Z. C. (2018). How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proc. of EMNLP2018*, pages 5010–5015.

Kawahara, D. and Kurohashi, S. (2006). A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proc. of HLT–NAACL2006*, pages 176–183.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL2014 (System Demonstrations)*, pages 55–60.

Morita, H., Kawahara, D., and Kurohashi, S. (2015). Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proc. of EMNLP2015*, pages 2292–2297.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP2016*, pages 2383–2392.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proc. of EMNLP2013*, pages 193–203.

Trischler, A., Ye, Z., Yuan, X., Bachman, P., Sordoni, A., and Suleman, K. (2016). Natural language comprehension with the EpiReader. In *Proc. of EMNLP2016*, pages 128–137.

Tsuchiya, M. (2018). Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proc. of LREC2018*, pages 1506–1511.