ATA2: A Question Answering System at NTCIR-13 QALab-3

Tao-Hsing Chang
National Kaohsiung University of Applied Sciences
415, Jiangong Rd., Sanmin Dist.
Kaohsiung 80778, Taiwan
changth@gm.kuas.edu.tw

Yu-Sheng Tsai
National Kaohsiung University of Applied Sciences
415, Jiangong Rd., Sanmin Dist.,
Kaohsiung 80778, Taiwan
1102108158@gm.kuas.edu.tw

Chih-Li Tsai
National Kaohsiung University of Applied Sciences
415, Jiangong Rd., Sanmin Dist.,
Kaohsiung 80778, Taiwan
1104108149@gm.kuas.edu.tw

Pei-Xuan Cai
National Kaohsiung University of Applied Sciences
415, Jiangong Rd., Sanmin Dist.,
Kaohsiung 80778, Taiwan
1104108124@gm.kuas.edu.tw

ABSTRACT
In 2016, we proposed a technique, called ASEE or ATA1, to automatically answer multiple-choice question. Multiple-choice questions refer to items where the best option is to be selected from the options provided. The core concept behind ATA1 is the idea that only the correct answer can become valid information and that there will be more valid information appearing on the Wikipedia. However, although such statistical method as ATA1 is very effective in processing multiple-choice questions, it cannot be used where the answer is not one of the options or on other types of questions, such as term questions that require an inference to find the answer. Therefore, this paper proposes a new tool for automatic answering called the ATA2. This tool will convert the content and Wikipedia page of the item into concept maps. A concept map is used to express the architecture of the knowledge. ATA2 compares the similarity between the concept maps of the item and source of knowledge to determine the answer. ATA2 can be applied to both of multiple-choice and term questions. This paper also shows the accuracy of ATA2 at QA-Lab 3.

Keywords
ATA2; concept map; question answering; entrance examination; term question; multiple-choice question.

Team Name
KUAS

Subtasks
Multiple-choice Question and Term Question

1. INTRODUCTION
In the field of information retrieval, a real-life problem is still unsolved: the use of machines to answer entrance examination items. Previously, we proposed a technique [2] to automatically answer multiple-choice question called ASEE or ATA1. Multiple-choice questions refer to items where the best option is to be selected from the options provided. The answering process of ATA1 is divided into four steps. The first step is to determine the type of question. ATA1 uses a rule-based module to determine the type of item as having different types affects the strategy to be used. Next, the Wikipedia search module finds out the Wikipedia page from which it can calculate whether an option is right or wrong. The third step is to use an evaluation formula to calculate the validity of each option on the Wikipedia page found in step two. The final step is to use an algorithm to compare the validity of each option to find the most likely answer.

The core concept behind ATA1 is the idea that only the correct answer can become valid information and that there will be more valid information appearing on the Wikipedia. However, although such statistical method as ATA1 is very effective in processing multiple-choice questions, it cannot be used where the answer is not one of the options or on other types of questions, such as term questions that require an inference to find the answer. Therefore, this paper proposes a new tool for automatic answering called the Automatic Test-Answering System II (ATA2). This tool will convert the content and Wikipedia page of the item into concept maps. A concept map is a widely used tool in science education, which expresses the architecture of the knowledge. ATA2 compares the similarity between the concept maps of the item and source of knowledge to determine the answer.

This paper will explain how ATA2 answers multiple-choice and term question items. The solutions of ATA2 can be applied to both of multiple-choice and term questions with some differences in parameters used in different steps. As such, this paper will present the techniques of ATA2 for term question items in section 3 and present adjustments that are required when using ATA2 for multiple-choice question type items. Section 4 will show the accuracy of ATA2 at QA-Lab 3 and, lastly, discuss the limitations and future works of ATA2.

2. RELATED WORKS
Novak and Canas [5] noted that a concept map is a tool to express knowledge. A focus question is first required to construct a concept map; the focus question is the subject that the constructor wishes to express and each concept is related to this subject. Afterwards, the concept map can be used to find concepts that are related to the focus question; a line between two concepts represents a relationship. A conjunction is used to account for the definition of this relationship, which then completes the concept map generated for the focus question.

In science education, teachers can use the concept maps drawn by students to evaluate the students' degree of understanding. Barroso and Crespillo [1] noted that concepts maps could be used to allow a person to better understand important concepts as it is a difficult task to help the average person understand complex subjects. The method of evaluating the students' concept maps is to compare their maps with the expert's concept map of the standard answer wherein the greater the similarity, the higher the

score. There are many methods for evaluating similarities, such as the method by [4] where the entirety, relationship, and structure of the concept map are used. However, these methods are mostly derived from the N-G rating method [6] and the closeness index rating method [3].

We consider items and Wikipedia pages to be collections of knowledge. By viewing each item as an expert's concept map and Wikipedia pages as a student's concept map, this means that the more similar the concept map of the item is to the concept map of the Wikipedia page, the more likely the answer will be located there. As far as we known, there are many studies for solving real-life university entrance exam questions [7][8], but the use of concept maps as a core concept has never been proposed. Also, the original definition of concept maps and similarity measurements must both be modified to conform to the requirements for automatic answering. Therefore, the proposed ATA2 in this paper has modified the definition of concept maps and proposed a new similarity measurement.

3. ANSWERING FOR TERM QUESTIONS

ATA2 divides the process of answering term questions into four steps. The first step is the formalization of items. Because some items contain two or more stems, ATA2 separates these items into several items with standard formats, allowing each item to have only a single stem. Next, ATA2 will select keywords from the items and form a query to be used by the search engine. The Google search engine will use this query to search and return the most similar Wikipedia page. ATA2 will treat these pages as ones containing the required information for answering the question. In the third step, ATA2 takes every sentence of these pages, converts them into concept maps, and then compares and labels them in comparison to the concept map of the item. Lastly, ATA2 will employ an algorithm to find the most likely answer from the sentences on the Wikipedia page. The following subsections contain the details of each of the above steps.

3.1 Item Formalization

The item of the term question consists of scenario and stem. Figure 1 is an example of an item where area A of Figure 1 represents the scenario while area B represents the stem. The scenario is used to suggest a related description of the item and the stem is the narrative required by the respondent to answer the question.

> The printing press played a major role in popularizing a German translation of the Bible in the early 1520s. (A)
>
> Write the name of the translator. (B)

Figure 1. Example of the standard item format

However, some items are different from Figure 1 as the item has two or more stems, as shown in Figure 2. In Figure 2, aside from A, there are also B and C which are two different stems; they all use the scenario A. When this type of item is parsed by using the Stanford Parser, B and C will each generate a parsing tree with a root node. Therefore, when the item is identified to have two individual parsing trees, they are considered to be two individual stems. As such, ATA2 involves the separation of the multiple stems followed by the addition of the original scenario to each stem, generating new items, as shown in Figure 3. By using these steps, no matter how many stems the item has, ATA2 can convert them into the standard format item with a single scenario paired with a single stem.

> During the Renaissance period, many churches were built in Italy with design features imitating aspects of ancient architecture. (A)
>
> Write in the name of a cathedral built in Florence. (B)
>
> Write in the name of the architect who designed its dome. (C)

Figure 2. Example of an item with multiple stems

3.2 Searching for Related Wikipedia Pages

In subsection 3.1, ATA2 parsed all the sentences contained in the item. Then ATA2 collects the nouns tagged as NN, NNS, NNP, and NNPS, verbs tagged as VB, VBD, and VBN but excluding "be," and words beginning with a capital letter in the item. As these words represent the important concepts, actions, and events of the item, they are combined to form a query, which is used on the Google search engine to find the most relevant pages for these keywords. Owing to the large number of web pages and the fact that the information provided may not necessarily be correct, the search is limited to Wikipedia pages.

> During the Renaissance period, many churches were built in Italy with design features imitating aspects of ancient architecture. Write in the name of a cathedral built in Florence. (A)
>
> During the Renaissance period, many churches were built in Italy with design features imitating aspects of ancient architecture. Write in the name of the architect who designed its dome. (B)

Figure 3. Result of the formalization of the item in Figure 2

After this, ATA2 will select the most related Wikipedia pages returned from the search as the basis for the next step in searching for the answer. The Google search engine will return the page hyperlinks and rank these hyperlinks according to their relevance to the query. Therefore, ATA2 only uses the top five Wikipedia pages because if ATA2 cannot find the answer among these pages, then the probability of finding the answer in other pages is even lower.

In addition, we observed in our experiments that if there are many nouns in the query, the use of only nouns in the query will sometimes find pages containing the answer more accurately than the method above because having more nouns helps to describe the meaning of the item in more detail. Therefore, ATA2 sets a threshold: when the number of nouns in the keywords exceeds the threshold, ATA2 will only use nouns to form the query to redo the search. ATA2 will collect three pages that are each returned by the first and second searches, respectively. These six pages will be the basis for the next step in searching for the answer.

3.3 Calculating the Validity of Candidates

ATA2 compares an item with every sentence in the Wikipedia pages selected by the method in subsection 3.2. It converts the item and a sentence in the pages into concept maps and calculates the degree of overlap between these two concept maps to determine whether the sentence contains the answer to the item.

The following is a hypothetical item that will be used to explain the process of constructing a concept map.

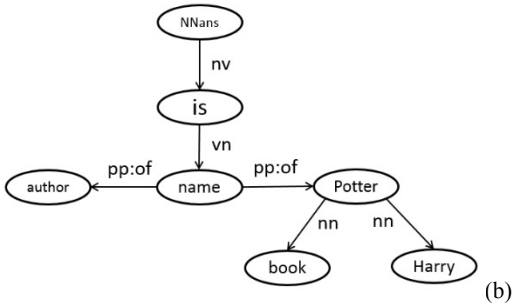
Write the author's name of the book Harry Potter.

To simplify the explanation, we have omitted the scenario of the item. Next, ATA2 uses a conversion module to convert the item into the following sentence:

NNans is the author's name of the book Harry Potter

where NNans represents the position of the answer to the question. If a sentence in the wiki-pages is the same as this item except the first word, then the first word of this sentence is the answer to this item. Figure 4(a) is the parsing tree of the converted item parsed by the Stanford Parser, As the most of sentences in Wikipedia pages are declarative sentences, their parsing trees can be generated directly by using parser.

```
(ROOT
 (S
  (NP (NNS NNans)
  (VP (VBZ is)
   (NP
    (NP
     (NP (DT the) (NN author) (POS 's)
     (NN name)
    (PP (IN of)
     (NP (DT the) (NN book) (NNP Harry) (NNP Potter)
 (. .)                                                (a)
```



(b)

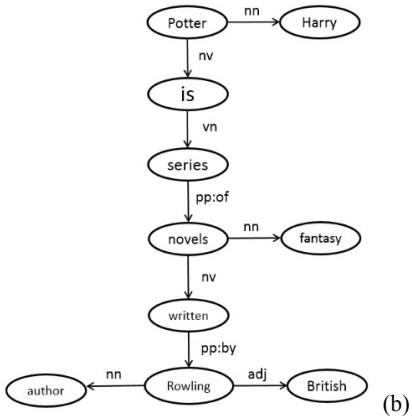
Figure 4. Parsing tree and concept map of an item

Using the parsing trees mentioned above, ATA2 convert the above item and every sentence in the Wikipedia pages into concept maps. Supposing that on the Wikipedia page there is a sentence as follows:

Harry Potter is a series of fantasy novels written by the British author Rowling.

The parsing tree of this sentence is shown in Figure 5(a). First, ATA2 selects words tagged with specific part-of-speech to be the node of the concept map. These specific parts of the speech include NNP, NN, V, JJ (adjective), CD (cardinal number), and PRP (personal pronoun). Next, ATA2 uses the relationship between the words in the parsing tree to establish the connections to the node above. Figures 4(b) and 5(b) is the concept map of Figures 4(a) and 5(a) after conversion, respectively. By using Figure 4(a), ATA2 can know that "Harry" and "Potter" are nouns that belong to the same level and, as such, will respectively establish the link "Potter->Harry" and mark the link with the tag "nn" to represent nouns on the same level. Also, "novels" and "written" in Figure 5(a) are a noun and verb that belong to the same level. ATA2 will point "novels" to "written" based on their sequence of appearance and mark both nouns that appear before the verb with the relationship "nv" while the relationship of "is," which points to "series," is "vn." Through the above method, ATA2 automatically converts the parsing tree into a concept map.

```
(ROOT
 (S
  (NP (NNP Harry) (NNP Potter)
  (VP (VBZ is)
   (NP
    (NP (DT a) (NN series)
    (PP (IN of)
     (NP
      (NP (NN fantasy) (NNS novels)
      (VP (VBN written)
       (PP (IN by)
        (NP (JJ British) (NN author) (NNP Rowling)
 (. .)                                                (a)
```



(b)

Figure 5. Parsing tree and concept map of a sentence in the Wikipedia pages

Concept maps for every sentence in the Wikipedia page can be generated in accordance with the above method. ATA2 will compare every concept map with that of the item. If two or more nodes of the concept map of the sentence are the same as that of the item, ATA2 will determine that this sentence is related to the item and consider it as one of the candidate sentences that contains the answer. If the opposite is true, the sentence is discarded. Finally, the candidate sentences gathered by ATA2 form a set of candidate sentences.

3.4 Deciding on the Answers

ATA2 assumes that the answer must be in one of the proper nouns in the set of candidate sentences and therefore will pick words with the tag NNP and NNPS (from now on referred to as candidate words) as candidate answers. ATA2 further assumes that the phrase with the item should appear in the same sentence as the answer word and these words should appear close to the location of the answer word. Based on the above assumptions, ATA2 designs an algorithm to calculate the probability of each candidate word to be the answer.

Figure 6 illustrates the calculation. Figure 6(a) is the concept map of an item, where Figures 6(b) and 6(c) are the concept maps of two candidate sentences collected, and that node B and node D in Figures 6(b) and (c) are candidate words. First, assuming that D is the answer and NNans is replaced with D is calculated, the score of the concept map of the item will be computed. This process is called the candidate word scoring.

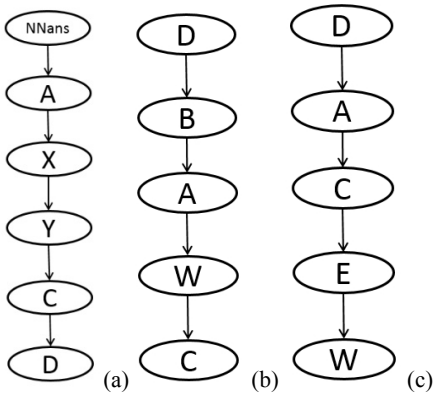


Figure 6. Example of the method to decide on the answer

This process first compares the next node of NNans, node A, to all the nodes in Figure 6(b) except for node D. When it is found that Figure 6(b) also contains node A, ATA2 will extract the chain between nodes NNans and node A in Figure 6(a) and the chain between nodes D and A in Figure 6(b). ATA2 will conduct a pairwise comparison of the nodes of both chains, shown in Figure 7. In Figure 7, the two chains contain five nodes. As is it assumed that D is the answer, NNans and D can be connected with a link, and that link is worth one point. Also, both chains also contain node A so both nodes can be connected by a link, which is also worth one point. Node B in Figure 7 has no the same node in the other chain, thus it only receives 0.8 points. Node A in Figure 6(a) receives the score 1*1*0.8 = 0.8.

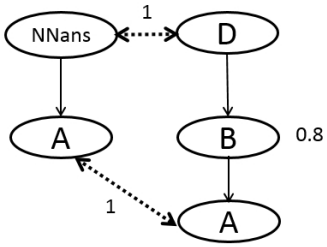


Figure 7. Calculation of scoring a candidate word

The candidate word scoring process will then continue to compare node X of Figure 6(a) with all the nodes of Figure 6(b). Because Figure 6(b) does not contain node X, node X is given zero points. Node Y of Figure 6(a) also gets zero points. Next, the score of node C of Figure 6(a) is calculated. ATA2 will extract the chain between nodes NNans and C in Figure 6(a) and the chain between nodes D and C in Figure 6(b). ATA2 will conduct a pairwise comparison of the nodes of both chains. In Figure 8, the two chains contain ten nodes and the six links between nodes worth one point each while the other four nodes without links obtain 0.8 points each. Therefore the total score of node C is 0.4096.

Using the above approach, every node of the item will receive a score. The candidate word will be scored with the next candidate sentence that contains this candidate word. If the node of the item receives a higher score, then the score of that node will be replaced by the higher score. Figure 9 illustrates an example. Node C respectively obtains a score of 0.4096 and 0.64 from Figure 6(b) and 6(c) if candidate word is node D. Therefore, the final score of node C is 0.64.

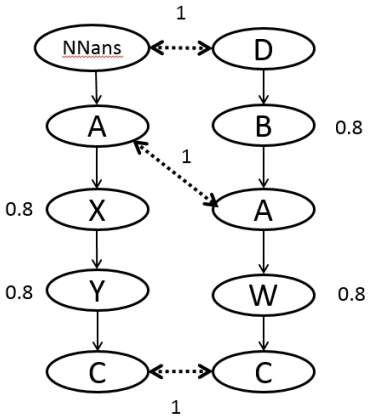


Figure 8. Calculation of scoring a candidate word

After the calculation mentioned above, every node of the item will have a score. The sum of these scores is the probability that candidate word D is the answer to the item. In Figure 6 for example, the probability that D is the answer to Figure 6(a) is 1.44. The same method can calculate the probabilities of all candidate words. Finally, the one with the highest probability is chosen as the answer. In the calculations described above, some words will be classified as stop words because they frequently appear in the Wikipedia pages. These stop words will not be candidate words. Also, a stop word does not form a connect between the concept maps of the item and candidate sentences.

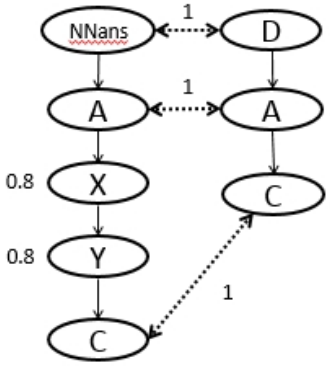


Figure 9. Example of how the probability of a candidate word being the answer is calculated

3.5 Solution for Multiple-Choice Questions

The steps for solving multiple-choice question items are roughly the same as that for term question items with four minor differences. First, as there are many types of items for multiple-choice questions, ATA2 uses a previous method [2] to categorize them into one of the five types: slot-filling items, single-word items, true-false items, combination items, and normal items. Second, there is no need for formalization because a multiple-choice question contains one stem only. Third, since the answer to a multiple-choice question is one of four options, ATA2 will create new queries by adding each option to the original query formed in subsection 3.2. It will increase the precision of searching through Wikipedia pages.

Finally, five type items apply different methods for determining the answer option. For single-word, combination, and slot-filling items, ATA2 will select the option that scored the highest in

subsection 3.3 as the answer. For normal items, where each option is a sentence, the sum of the scores of each proper noun in a option is the score of the option. The option with the highest score is the final answer.

For true-false items, the process for determining the answer option is more complicated than that for other types because true-false items do not have stems and instead require the examinee to determine whether the scenario is right or wrong. Therefore, ATA2 will start by finding the first proper noun in each sentence where correctness needs to be determined. Then the proper noun will be substituted by the NNans mark mentioned in subsection 3.3, which will convert the sentence into a stem. Next, the score of the noun can be computed by using the methods described in subsections 3.2 to 3.4. Afterwards, the second proper noun of the sentence is found, and the same method is repeated to obtain a score. By repeating this, the score of every proper noun in the sentence can be calculated. Finally, the sum of these scores is the score of the sentence. If the score of the sentence exceeds a threshold, ATA2 will consider the sentence to be "true."

Apart from true-false items, if items go through the same calculation previously mentioned and result in the score of any option is not different from that of other options, ATA2 will redo the previously mentioned process with a new query. The query only consists of the nouns and verbs in the option. If it is still not possible to determine which option is the right one, then a statistical method [2] will be used.

4. EXPERIMENTAL RESULTS

The collections used in the experiment were provided by the NTCIR-13 QALab-3 [9]; they originate from World History Exam B of the National Center for University Admissions in Japan. The training data of the collection are collected from the exams in 2012 and 2013 while the test data are collected from the exam in 2014. There were a total of 72 multiple-choice questions in the training data and 36 of them in the test data. There were a total of 68 term questions in the training data and 77 of them in the test data. The data was in the XML format. As the training and test data contain items that are images or items that ask about the chronology of events, these items were classified by ATA2 as unidentified items.

Table 1. Accuracy in answering multiple-choice question type test items

Item types	# of test questions	# of correct answers	Accuracy
Normal	11	5	0.45
Slot-filling	1	0	N/A
Single-word,	4	1	0.25
Combination	9	8	0.89
True-false	7	2	0.29
unidentified	4	N/A	N/A
All items	36	16	0.44

Table 1 shows the accuracy of using ATA2 to answer multiple-choice questions from the test collection. Table 1 also shows the experimental results of each item type. It is worth noting that there is a slot-filling item in the 2014 test data that was not answered, which is why the accuracy for slot-filling items is zero. As there are few slot-filling items, Table 1 cannot be used to directly infer the accuracy of ATA2 in handling this type of item.

Table 2 shows the accuracy of ATA2 for answering term questions. Also, some items contain two or more question. The items to which ATA2 answers partial questions correctly is classified into "partially correct" in Table 2. In addition, some words ATA2 answers are very similar to answers. For example, the answer of an item is "Zionism" while that of ATA2 is "Zion". These items are also grouped into "partially correct". In Table 2, although an accuracy of 0.14 is quite a low figure, it is not due to the poor accuracy when concept maps are used to answer questions; rather it is because approximately 65% of all items do not have answers in Wikipedia pages, which is why ATA2 was unable to answer these items. Therefore, the concept map answering method should be considered an effective method.

Table 2. Accuracy in answering term question type test items

	# of Test Items	Completely Correct	Partially Correct	Accuracy
Processed	73	6	4	0.14
Unidentified	4	N/A	N/A	N/A
All Items	77	6	4	0.13

5. DISCUSSION AND FUTURE WORKS

Experimental results shows that it is feasible and effective to use the concept map-based model of automatically answering examination items. However, it is still not possible to correctly answer some items, which is mainly due to three reasons. First, the only source of knowledge for this method is Wikipedia. If the content that answers the question is not on Wikipedia, then it is impossible to arrive at the answer. Second, some of the concepts in the concept map generated by ATA2 are not important, leading to longer candidate sentences that may not obtain a higher score. Although we have tried using the list of stop words to reduce unnecessary concepts, it is not easy to determine whether some concepts are stop words. Third, many parts of the score calculation require thresholds and parameters that are derived from experimental experience. Whether these values are the optimal ones remains to be verified.

Therefore, ATA2 should perform better if the three problems above are addressed. This question-answering method is based on the knowledge structure comparison; it could be an effective method to address problems of extraction and generation in processing knowledge. In the future, the application of this method to other types of items, such as essay questions, can be considered.

6. ACKNOWLEDGMENTS

This work is supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grant MOST 104-2511-S-151-001-MY3

7. REFERENCES

[1] Barroso, C. and Crespillo, R. 2008. Concept maps: Tools For Understanding Complex Problems. In Proceedings of the 3rd International Conference on Concept Mapping.

[2] Chang, T. H. and Tsai, Y. S. 2016. ASEE: An Automated Question Answering System for World History Exams. In Proceedings of the 12th Conference on Evaluation of Information Access Technologies, NTCIR 12, Tokyo, Japan, 445-450,

[3] Goldsmith, T. E., Johnson, P. J., and. Acton, W. H. 1991. Assessing structural knowledge. Journal of Educational Psychology, 83, 88-96.

[4] McClure, J. R., Sonak, B., and Suen, H. K. 1999. Concept Map Assessment of Classroom Learning: Reliability, Validity, and Logistical Practicality. Journal of Research in Science Teaching, 36, 4, 475–492.

[5] Novak, J. D. and Canas, A. J. 2007. Theoretical Origins of Concept Maps, How to Construct Them, and Uses in Education. Florida Institute for Human and Machine Cognition, 3, 1, 29-42.

[6] Novak, J. D. and Gowin, D. B. 1984. Learning how to learn. New York: Cambridge University Press.

[7] Shibuki, H., Sakamoto, K. Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T., and Kando, N. 2016. Overview of the NTCIR-12 QA Lab-2 Task. In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, 392-408, Tokyo, Japan.

[8] Shibuki, H., Sakamoto, K., Kano, Y., Mitamura, T., Ishioroshi, M., Itakura, K. Y., Wang, D., Mori, T., and Kando, N. 2014. Overview of the NTCIR-11 QA-Lab Task. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, 518-529, Tokyo, Japan.

[9] Shibuki, H., Sakamoto, K., Ishioroshi, M., Kano, Y., Mitamura, T., Mori, T., and Kando, N. 2017. Overview of the NTCIR-13 QA Lab-3 Task. Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan