KitAi-QA: a Question Answering System for NTCIR-12 QALab-2

Shuji Fukamachi Kazutaka Shimada

Department of Artificial Intelligence
Kyushu Institute of Technology
{s_fukamachi, shimada}@pluto.ai.kyutech.ac.jp

ABSTRACT

This paper describes a question answering system for NTCIR-12 QALab-2. The task that we participated in is the Japanese task about National Center Test and Mock exams. Our method consists of two stages; a scoring method and answer selection methods for four question types. The scoring is to detect the evidence for the next process, namely answer selection, from textbooks. We also focus on conflict detection and event detection for the answer selection of the True-or-False type question. For other questions, Factoid, Slot-filling and Unique Time, our method judges or extracts the answer from the passage retrieved by the scoring method. The accuracy of our method on the formal run was moderate. However, the result of our method sometimes boosted up other system results on the combination run. The result shows the effectiveness of our method.

Team Name

KitAi / Kyushu Institute of Technology (Department of Artificial Intelligence)

Subtasks

National Center Test and Mock exams (Japanese)

Keywords

Scoring, World history ontology

1. INTRODUCTION

This paper describes our question answering system, KitAi[1]-QA, for NTCIR-12 QALab-2 [5]. Our method identifies the question type of each question on the basis of some surface expression rules. After question type identification, our method retrieves the most related topic (a unit in textbooks). For the detection of the related topic, we apply a scoring approach [1]. We compute an importance value of each word in textbooks, and then detect the related topic by using queries and the importance value. Then, our method select the answer in each question type; True-or-False, Factoid, Slot-filling and Unique Time. We focus on conflict detection and event detection for the True-or-False type question. The conflict detection is to recognize conflict between a query and the related topic. The event detection labels a true or a conflict tag to a choice on the basis of verb predicate structures and the world history ontology developed by [2]. For other questions, Factoid, Slot-filling and Unique Time, our method judges or extracts the answer from the passage retrieved by the scoring method. Figure 1 shows the outline of our system. In other words, our method consists of three stages; question type identification, a scoring method and answer selection methods for four question types.

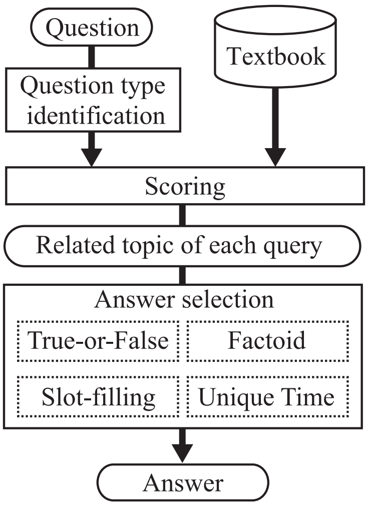


Figure 1: The outline of our method.

In the next section, we describe the question type identification. Next, we describe the scoring method for retrieving a passage, namely a related topic for the answer and answer selection methods for True-or-False, Factoid, Slot-filling and Unique Time questions. Then, we discuss our experimental results in Section 5. Finally, we conclude our method in Section 6.

2. QUESTION TYPE IDENTIFICATION

In our method, we handle five types of question; True-or-False, Factoid, Slot-filling, Unique Image and Unique Time. The identification of the question type is based on rules with some surface expressions. The rules are as follows:

True-or-False: The question contains one of the following patterns;

- “[述べた文|述べた次の文]” and “[正しい|適当|誤っている|誤りを含む] もの”)

[1] Short of *K*yushu *I*nstitute of *T*echnology (Department of *A*rtificial *I*ntelligence). The English meaning is “expectation.”

- “[事柄|事績|出来事] として” and ”[正しい|誤]”

Slot-filling: The question contains “空欄” or “[入る|入れる]語”

Unique Image: The question contains “写真”, “図”, “グラフ” or “絵の中の”

Unique Time: The question contains “年代” and “時期” with “配列” for Unique Time Reordering or the question contains “[期|時期] として” for Unique What Time

Factoid: Other than the above

We identify the five question types. However, we handle four question types, True-or-False, Factoid, Slot-filling and Unique Time, in the following process. In other words, we ignore questions about Unique Image[2].

3. SCORING FOR RELATED TOPIC

In this section, we explain our scoring method. The purpose of the scoring is to detect the evidence for the next process, namely answer selection, from textbooks. In our method, a passage extracted by using the scoring process is called “related topic.”

3.1 Query for scoring

To detect a related topic is to retrieve an important passage from the textbooks. Therefore, we need queries for the retrieval process. The queries depend on the question type of each question.

True-or-False: We use four query sets from four choices in the question. A query set consists of some query words. In addition, if the question contains the phrase “この XXX について (about this XXX)”, we add words from sentences including “XXX” in the description part related with the question. We use named entities as query words[3].

Slot-filling: We use one query set from sentences with slots. We use all nouns as query words.

Unique Time: For Unique Time Reordering, we use some query sets from choices in the question. We use named entities for this question type as query words. For Unique What Time, we use one query set from the question. We use all nouns for this question type as query words.

Factoid: We use one query set from the question. In addition, if the question contains the phrase “この XXX について (about this XXX)”, we add words from sentences including “XXX” in the description part related with the question. We use all nouns for this question type as query words.

3.2 Keyword extraction

The textbooks have been annotated with topics for each paragraph manually. To extract the most related topic from the textbooks for the answer selection, we need to compute an importance measure of each word in each topic.

First we extract nouns and the dependency relations from each sentence in the textbooks by using a morphological analysis tool Juman[4] and a dependency parser KNP[5]. For more accurate analysis, we apply instances of the world history ontology [2] and entities annotated by QALab organizers into the dictionary of Juman. Here we introduce some rules for notation fluctuation problems.

Deletion We prepare some prefix and suffix patterns for this process. For example, 大統領 (President) is a suffix pattern. We handle a word without a suffix pattern as different notation of the original expression. For example, we generate オバマ (Obama) from the expression オバマ大統領 (President Obama[6]), and then we use {オバマ大統領, オバマ} as a word list.

Combination In Japanese, there are many Katakana expressions in the textbooks about the the world history. These words have an important role for the answer selection. We generate several combinations from a phrase with Katakana expressions. For example, we generate USA, Obama and USA-Obama from ‘’USA President Obama.”

Expansion Some words in the world history ontology contain another description; e.g., EU and European Union. By using this knowledge, we expand the dependency of the original relation. For example, if there is a pattern ‘’The EU is a politico-economic union”, we generate a new pattern “The European Union is a politico-economic union.”

By using these rules, the coverage of words is improved.

Then, we compute an importance value of each keyword candidate w_i. The value is based on *idf* in terms of each topic in the textbooks.

$$Imp(w_i) = \log \frac{AllTopics}{NumTopic(w_i)} \quad (1)$$

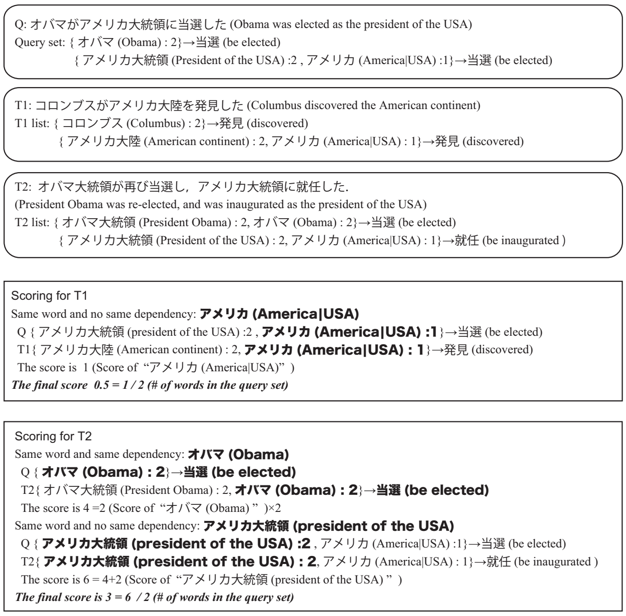
where $AllTopics$ is the number of topics in the textbooks. $NumTopic(w_i)$ is the number of topics that contain a word w_i. w_i is a word after application of the rules about different notation problems. Assume that a topic contains the word “United State of America” and another topic contains the word “President of America”. In this situation, the importance value of the word “America” is smaller than “United State of America” and “President of America” because the word “America” appears in both topics.

3.3 Related topic extraction

By using keywords and the importance value *Imp* computed in Section 3.2, we compute a score between a query set and each topic in the textbooks. First, we set the score to zero. Next, we retrieve topics in the textbook with a query in the query set. Here we handle a word list, such as {オバマ大統領, オバマ} in Section 3.2, for the query. In a similar manner, we handle a word list for each word in each topic. If a word in the word list of a query matches with a word in the word list of a topic and the dependency of the query word matches with that of the topic word, the importance value of the word is multiplied by 2, and then we add the

[2] Properly speaking, we select the answer about this question type randomly in the answer selection process.
[3] For combination True-or-False questions, we use all nouns.

[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[6] In Japanese, the word “President (大統領)” is a suffix.



Q: オバマがアメリカ大統領に当選した (Obama was elected as the president of the USA)
Query set: { オバマ (Obama) : 2}→当選 (be elected)
{ アメリカ大統領 (President of the USA) :2 , アメリカ (America|USA) :1}→当選 (be elected)

T1: コロンブスがアメリカ大陸を発見した (Columbus discovered the American continent)
T1 list: { コロンブス (Columbus) : 2}→発見 (discovered)
{ アメリカ大陸 (American continent) : 2, アメリカ (America|USA) : 1}→発見 (discovered)

T2: オバマ大統領が再び当選し，アメリカ大統領に就任した.
(President Obama was re-elected, and was inaugurated as the president of the USA)
T2 list: { オバマ大統領 (President Obama) : 2, オバマ (Obama) : 2}→当選 (be elected)
{ アメリカ大統領 (President of the USA) : 2, アメリカ (America|USA) : 1}→就任 (be inaugurated)

Scoring for T1
Same word and no same dependency: **アメリカ (America|USA)**
Q { アメリカ大統領 (president of the USA) :2 , **アメリカ (America|USA) :1**}→当選 (be elected)
T1{ アメリカ大陸 (American continent) : 2, **アメリカ (America|USA) : 1**}→発見 (discovered)
The score is 1 (Score of "アメリカ (America|USA)")
The final score 0.5 = 1 / 2 (# of words in the query set)

Scoring for T2
Same word and same dependency: **オバマ (Obama)**
Q { **オバマ (Obama) : 2**}→**当選 (be elected)**
T2{ オバマ大統領 (President Obama) : 2, **オバマ (Obama) : 2**}→**当選 (be elected)**
The score is 4 =2 (Score of "オバマ (Obama) ")×2
Same word and no same dependency: **アメリカ大統領 (president of the USA)**
Q { **アメリカ大統領 (president of the USA) :2** , アメリカ (America|USA) :1}→当選 (be elected)
T2{ **アメリカ大統領 (president of the USA) : 2**, アメリカ (America|USA) : 1}→就任 (be inaugurated)
The score is 6 = 4+2 (Score of "アメリカ大統領 (president of the USA) ")
The final score is 3 = 6 / 2 (# of words in the query set)

Figure 2: An example of the scoring process.

value to the score. If there is the same word and not the same dependency between the query and the topic, we add the importance value of the word to the score. Our method performs this process for all words in a query set. Then, we divide the score by the number of words in the query set. We compute the score of all topics, and then extract the topic with the maximum score as the related topic of the query set. Figure 2 shows an example of this scoring method. In the figure, the curly brace ({ }) denotes a word list and the arrow → denotes a dependency relation. The value of each word, e.g., オバマ (Obama):2, is Imp. If a word list contains some values, such as {アメリカ大統領 (president of the USA) : 2, アメリカ (America|USA) : 1} in T2, we use the maximum value in the list. In this example, T2 is selected as the related topic of the query set.

4. ANSWER SELECTION

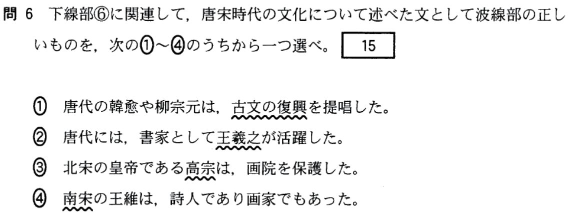
The answer selection consists of four processes. They are based on the question types.

4.1 True-or-False

The true-or-false question is a fact validation task of choices. For the true-or-false question, our method contains two types of preprocessing; conflict detection and event detection. We judge the final answer of the fact validation task by using the results of the preprocessing and the score in Section 3.

4.1.1 *Conflict detection*

The first process is conflict detection. The purpose of this process is to recognize conflict in choices in each question. If there is a conflict of a choice, our method labels a conflict tag to the choice. The conflict detection consists of two types of conflict; time conflict and wave-line conflict.



問 6 下線部⑥に関連して，唐宋時代の文化について述べた文として波線部の正しいものを，次の①～④のうちから一つ選べ。 15

① 唐代の韓愈や柳宗元は，古文の復興を提唱した。
② 唐代には，書家として王羲之が活躍した。
③ 北宋の皇帝である高宗は，画院を保護した。
④ 南宋の王維は，詩人であり画家でもあった。

Figure 3: An example of the wave-line question.

The conflict on time information denotes a crucial point on the true-or-false question. Some researchers have focused on this point [3, 4, 7]. In our method, time information is expressed as follows; "1901-2000" for "twentieth century" and "2000-2000" for 2000th year. We also estimate time information about events with Wikipedia; e.g., 1368-1644 for " 明の時代 (the Ming dynasty in China)" and 1939-1945 for " 第二次世界大戦 (World War II)". We handle "named entity + time expression" as the target pattern of the time estimation. For example, "第二次世界大戦 以降 (after WWII)" and "20 世紀 前半 (the early 20th century)". In addition, time information depends on the time expressions. For example, 1939-1945 for "第二次世界大戦 中 (during WWII)" and 1945-now for "第二次世界大戦 以降 (after WWII)". If there is a conflict between the time information in the choice and sentences in the related topic, our method rejects the choice as a false event. If the related topic contains some sentences of different time information, we regard the most early year and the most late year in sentences as the begin and end time information.

The second conflict relates with a words with a wavy line in choices. Figure 3 shows an example of a question and the choices. In this situation, words except a word with a wavy line are always true. Hence, if there is no mention about the wave-line word in the related topic, we can label a conflict tag to the choice. Properly speaking, we extract the related topic without the wave-line word in Section 3.3, and then check whether the wave-line word exists in the related topic or not. If the wave-line word does not exist in the related topic, the choice is false.

4.1.2 *Event detection*

The task in QALab relates with history questions. Therefore, choices on each question are past events in the world history. Okita and Liu [4] have proposed a method with verb predicate structures and knowledge for the same task. We also focus on verb predicate structures with world history ontology for event detection. We generate a verb predicate structure of each sentence by using KNP. For example, our method obtains the verb predicate structure[7] "win(Agent: XX, Object: YY, Situation: ZZ)" from a sentence "XX defeated YY in the ZZ battle." We also apply some rules into the process;

- passive: we exchange ga-case and wo-case if the verb in a sentence is passive. For example, if a sentence is "毒ガスが塹壕戦で使われた (The poison gas was used in the trench warfare)" , we obtain the verb predicate structure "use(Object: poison gas, Situation: trench warfare)".
- is-a: we create another structure if a sentence has "is-a"

[7] Assume that the word "defeat" and "win" are the same event in the ontology.

relation, “XX は YY である (XX is YY)”. For example, “八旗は順治帝が創設した軍隊である (Hakki was the army that Junchitei set up,) has two structures; “setUp(Agent: Junchitei, Object: Army)” as the original structure and “setUp(Agent: Junchitei, Object: Hakki)” as the is-a relation.

- different notation: We apply the rules in Section 3.2 to the predicate structure. For example, if “win(Agent: President Obama)” is generated from a sentence, we also generate “win(Agent: Obama)”.

Our method identifies a relation between the verb and an event in the ontology. We use the definition of simple events in [2]. Here we regard all events in the textbooks as the true events. If there is a conflict between the event in a choice and any events in the textbooks, e.g., win(XX, YY) and lose(XX, YY), we label a conflict tag to the choice. On the other hand, if there are win(XX, YY) and win(XX, YY) or win(XX, YY) and lose(YY, XX), we label a true tag to the choice.

4.1.3 Final decision

After conflict detection and event detection, namely labeling about true and conflict tags, we judge the true-or-false answer of the questions by using the score in Section 3.3 and the tags in Section 4.1.1 and 4.1.2.

The final decision is based on the ranking of scores of each choice. We use the scores described in Section 3.3 as the scores in this process. If the choice contains a negation, such as “なかった”, the score is inverted. First, if a choice has a true tag, we alter the score of the choice to more than the 1st rank score. On the other hand, if a choice has a conflict tag, we alter the score of the choice to less than the lowest rank score. Then, if the question is the selection of the correct event, our method extracts the 1st rank choice. In a similar manner, if the question is the selection of the incorrect event, our method extracts the the lowest rank choice. In this situation, if the difference between the scores of the 1st rank and 2nd rank (or the lowest and 2nd to lowest) is small[8], we re-compute scores of all topics by using words in the description part. Then, we select the correct (or incorrect) choice on the basis of the re-computed score.

Some true-or-false questions require a combination of true or false of choices. In this situation, we need to judge the true-or-false of all choices because the score just denotes a true-or-false likelihood, that is not true-or-false itself. Therefore, we judge the flag of each choice with a threshold. In this paper, we set it to 1.09. We treat that a choice is true if the score is the threshold or more. Otherwise, it is false.

4.2 Factoid

The factoid question is to select the fact (or non-factt) word from choices. If the question is the selection of the fact, we select the word with the maximum frequency in the related topic detected in Section 3. If the question is the selection of the non-fact, we select the word with the minimum frequency in the related topic. If the frequency of any choices is zero, we re-compute the score about underlined words in the description part and then select the most suitable word in a similar way. If the question is the selection of the correct combination of words, we select the combination with the maximum frequency in the related topic detected in Section 3.

4.3 Slot-filling

The slot-filling question is to select the correct word (or combination) from choices. If the question is the selection of the correct word, we select the word with the maximum frequency in the related topic detected in Section 3. If the question is the selection of the correct combination of words, we select the combination with the maximum frequency in the related topic detected in Section 3.

4.4 Unique Time

The unique time question is to select the correct sequence of events from choices. For the time sequence detection, we use the same method in Section 4.1.1. We estimate the time of each event by using the detected time information. If the method can not capture the time information from sentence in the question, it estimate time information from events in the related topic. On the basis of the estimated time information, we select the suitable sequence from the choices.

5. EXPERIMENTAL RESULT

We participated all phase, namely phase-1, phase-2 and phase-3. We focused on National Center Test and Mock Exams. We constructed some systems (Priority 1, 2 and 3) in each phase. For the phase 1 and 2, the priority 1 denotes a system based on the scoring and conflict detection and the priority 2 denotes a system using all features in this paper. For the phase 3, the priority 1 denotes a system using all features in this paper. The priority 2 and the priority 3 denote a system based on the scoring and conflict detection and a system based on the scoring and event detection, respectively.

The results on the formal run are shown in Table 1. Our method was moderate. On the other hand, the result of our method sometimes boosted up other system results. Table 2 shows a part of the combination run. The correct rate of the two systems in the table drastically were improved by using our result (0.25 vs. 0.36 for Team Forst and 0.11 vs. 0.33 for Team IMTKU). These results show the effectiveness of our method.

We introduced some rules for different notation problems in Section 3.2. However, it was insufficient. Moreover, the different notation was the main reason of mistakes in the scoring method. Therefore, handling the different notations correctly is one of the most important future work.

The conflict detection was effective in the experiment. However, the number of target instances for the conflict detection was not very great. The event detection was not effective in the experiment. The reason was that our predicate structures were too simple. They did not capture context information and paraphrasing. In addition, handling causal relation and time relation between events [6] is an important future work.

6. CONCLUSIONS

This paper describes a question answering system based on a scoring method for NTCIR-12 QALab-2. For the TrueorFalse question, we introduced conflict detection based on

[8] The difference threshold is 1.0 in this paper.

Pahse1: Center Test			
Priority	Score	Correct Rate	Rank
1	29	0.27	18/23
2	29	0.27	19/23
Pahse1: Besesse Test			
2	41	0.42	1/7
1	38	0.39	2/7
Pahse1: Yozemi Test (1)			
1	24	0.25	2/7
2	24	0.25	2/7
Pahse1: Yozemi Test (2)			
1	28	0.28	7/7
Pahse2: Besesse Test			
2	64	0.37	11/18
1	62	0.35	15/18
Pahse3: Center Test			
1	31	0.31	18/32
2	31	0.31	18/32
3	31	0.31	18/32
Pahse3: Besesse Test			
2	40	0.31	6/12
1	37	0.29	9/12
3	34	0.27	10/12
Pahse3: Yozemi Test (1)			
2	39	0.39	3/9
1	38	0.39	4/9
3	38	0.39	4/9
Pahse3: Yozemi Test (2)			
2	30	0.28	6/9
3	30	0.28	6/9
1	24	0.22	9/9

Table 1: The formal run result.

Pahse1: Besesse Test				
Priority	Comb	Priority	Score	Correct Rate
Forst	-	2	26	0.25
Forst	KitAi	2	35	0.36
Pahse3: Center Test				
Priority	Comb	Priority	Score	Correct Rate
IMTKU	-	1	12	0.11
IMTKU	KitAi	1	34	0.33

Table 2: The combination run result.

time information and event detection based on world history ontology. For other questions, Factoid, Slot-filling and Unique Time, our method extracted the answer from the passage retrieved by the scoring method.

The accuracy of our method was moderate. However, the result of our method sometimes boosted up other system results on the combination run. The result shows the effectiveness of our method.

7. REFERENCES

[1] Y. Kano. Solving history exam by keyword distribution: Kjp system at ntcir-11 qalab task. In *In Proceedings of the 11th NTCIR Conference*, 2014.

[2] A. Kawazoe, Y. Miyao, T. Matsuzaki, H. Yokono, and N. Arai. World history ontology for reasoning truth/falsehood of sentences: Event classification to fill in the gaps between knowledge resources and natural language texts. In *New Frontiers in Artificial Intelligence (JSAI-isAI 2013 Workshops), Lecture Notes in Computer Science 8417*, pages 42–50, 2014.

[3] Y. Kimura, F. Ashihara, A. Jordan, K. Takamaru, Y. Uchida, H. Ototake, H. Shibuki, M. Ptaszynski, R. Rzepka, F. Masui, and K. Araki. Using time periods comparison for eliminating chronological discrepancies between question and answer candidates at qalab ntcir11 task. In *In Proceedings of the 11th NTCIR Conference*, 2014.

[4] T. Okita and Q. Liu. The question answering system of dcumt in ntcir-11 qa lab. In *In Proceedings of the 11th NTCIR Conference*, 2014.

[5] H. Shibuki, K. Sakamoto, M. Ishioroshi, A. Fujita, Y. Kano, T. Mitamura, T. Mori, and N. Kando. Task overview for ntcir-12 qa lab-2. In *In Proceeding of the 12th NTCIR Conference*, 2016.

[6] H. Takagi and K. Shimada. Time sequence estimation using semantics and latent temporal value of events. In *In Proceedings of the 22nd Annual Meeting of The Association for Natural Language Processing (in Japanese)*, 2016.

[7] D. Wang, L. Boytsov, J. Araki, A. Patel, J. Gee, Z. Liu, E. Nyberg, and T. Mitamura. Cmu multiple-choice question answering system at ntcir-11 qa-lab. In *In Proceedings of the 11th NTCIR Conference*, 2014.