SML Question-Answering System for World History Essay and Multiple-choice Exams at NTCIR-12 QA Lab-2

Takuma Takada
Nagoya University
takuma_t@nuee.nagoya-u.ac.jp

Takuya Imagawa
Nagoya University
t_imagaw@nuee.nagoya-u.ac.jp

Takuya Matsuzaki
Nagoya University
matuzaki@nuee.nagoya-u.ac.jp

Satoshi Sato
Nagoya University
ssato@nuee.nagoya-u.ac.jp

ABSTRACT

This paper describes SML team's approach to automatically answering world history exam questions at NTCIR-12 QALab. We challenged to answer both the multiple-choice questions in the national center tests and the essay-type questions in the secondary exams. Our system answers the questions by searching based on surface similarity. We explored several methods to enhance the system with domain-specific knowledge such as dictionaries of synonyms and temporal information.

Keywords

Question-Answering, Surface Similarity,Essay Generation, Compression, Category Prediction

Team Name

SML

Subtasks

Japanese

1. INTRODUCTION

Automatically solving Japanese university exams is one of the most challenging tasks for question-answering technologies [1]. By challenging these exams, we can measure the degree of achievement of the current question answering technology and identify the remaining issues from the viewpoint of the diversity of Japanese university exams and the variety of knowledge required there.

Especially, the essay-type questions of world history are different from the question types that previous researches have focused on, such as the factoid-type (e.g., Who is the founder of Edo shogunate?) and the why-type (e.g., Why is the sky blue ?). The questions of the essay-type exam of world history include, for instance, those on the process of historical events (e.g., Summarize within 60 characters the formation process of the police in ancient Greece.) and some of them do not clearly specify what aspect of the subject is supposed to be answered (e.g., Answer within 60 characters about the Cities Alliance, which was formed in northern Italy.). As a first step toward these new challenges, we developed an automatic answering system for the world history essay-type questions that generates an answer as an extractive summary of the textbook.

Meanwhile, most of the multiple-type questions are the fact-validation (FV) type, which we can answer by recognizing textual entailment between textbooks and the questions. Hattori et al. developed textual entailment recognition system for world history multiple-choice exams based on surface-similarity[2]. Their system scores the sentences in a textbook based on their similarity to a choice in a question. If it is higher than a threshold, their system further tests whether the choice is true or false by the overriding rules based on temporal information and named entities. We improved this system to challenge NTCIR-12 QALab.

The rest of the paper is organized as follows. Section 2 and 3 describe the detail of the answering system for the world history essay-type questions. Section 4 describes the factoid answering system. Section 5 describes the answering system for multiple-choice questions system. Section 6 describes the evaluation results of each system. Section 7 provides an analysis and discuss evaluation results on the world history essay-type questions.

2. WORLD HISTORY SHORT ESSAY-TYPE QUESTION AUTOMATIC ANSWERING SYSTEM

Figure 1 shows examples of world history essay-type questions. We found that we can answer most questions by extracting sentences from the textbook and combining them. We thus developed a system that generates an answer by extracting sentences from textbooks based on the degree of similarity to question sentences. As for the similarity score, we improved the score based on superficial similarity proposed by Hattori et al [2]. Of course, for the world history essay questions, which include a mixture of various question types, we do not believe that we can solve those questions by only a summary method based on surface similarity. The aim is, through the evaluation and analysis of the output of the system, to obtain insights towards a system based on a deeper analysis of the question sentences.

2.1 System Overview

Figure 2 shows the system block diagram. This system consists of four blocks: Search, Temporal relations labeling,

キリスト教徒がローマ皇帝に迫害された理由を
60字以内で説明しなさい．(2013年度東大)

5世紀におけるフン族の最盛期とその後について，
60字以内で説明しなさい．(2012年度東大)

Figure 1: World history short essay-type questions

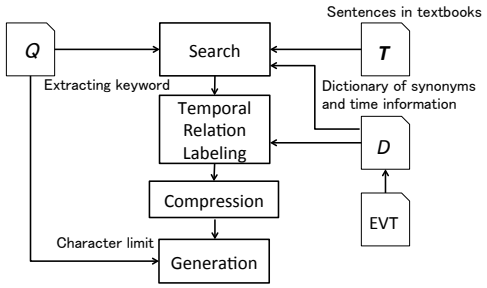

Figure 2: System block diagram

Compression, and Generation. In the experiment, we used four textbooks provided by the task organizers [3] as the source of sentences that make up the answer.

2.1.1 Search

The search block scores each sentence of the textbooks on the basis of the surface similarity between the textbook and question sentences. It outputs a list of the sentences in the descending order of their scores. The score is defined as follows:

$$score = \frac{\sum_{x \in \boldsymbol{N^1}} \{\min(f(x,t), f(x,q)) \cdot a(x) + 0.3 \cdot c\}}{\sum_{x \in \boldsymbol{N^1}} f(x,q)} \quad (1)$$

where t is a sentence in the textbooks, q is the sentence in the question, and $f(x,t)$ is the number of times the element x of the set $\boldsymbol{N^1}$ appears in t. As the set $\boldsymbol{N^1}$, we used noun unigrams. $a(x)$ is a weighting function defined as follows:

$$a(x) = \begin{cases} 1 & \text{if } x \in \text{Headword of the glossary} \\ 0.7 & \text{else if } x \in \text{Named entities} \\ 0.5 & \text{else if } x \in \text{Wikipedia page title} \\ 0.1 & \text{otherwise.} \end{cases}$$

Variable c indicates the position of x in q, counted by the number of the words that precede x in q. This is based on the observation that a noun that appears later in the question is often an important keyword and hence it should contribute significantly to the score. Furthermore, we made a dictionary of synonyms of world history terms that is extracted from a historical event ontology (EVT) [3]. Before the score calculation, nouns in the synonym dictionary were rewritten to their canonical forms defined in the dictionary.

2.1.2 Temporal relation labeling

When a sentence in the textbooks and the question mention to some historical events that happened at different time points, the sentence in the textbooks is less likely to be a sentence to include in the correct answer. We hence extract temporal information from the textbook sentence and the question (e.g., "from 1800 to 1843" → 1800-1843), compare them, and judge whether or not they match. When the sentences include temporal expressions (e.g., "in the 13th century","in 1745"), we extract them. Otherwise, we extract the temporal information by using a time dictionary made by extracting the mapping from historical events and persons to the times they occur or exist from EVT. The relationships between the temporal information of the problem and textbook sentences were classified as follows:

Match: Temporal information of the question matches to that of the textbook sentence

Unknown: No temporal information extracted either from the the textbook or the question

Mismatch: Temporal information of the question mismatches to that of the textbook sentence

2.1.3 Compression

We may adopt a simple method that chooses the sentences that fit within the character limit and have top scores to generate the essay. However there is a risk of choosing a sentence scored very low only because it fits within a very short character limit such as 30 characters. A low-scored sentence is expected to be mostly incorrect. We thus worked for resolution by fitting a high scored sentences within the character limit. Our current system solves this problem by compressing the sentence when the best-scored sentence that fits within the character limit and has a lower score than a threshold. The procedure of the compression is:

1. We compare the highest score of the sentences that fit within the character limit with the threshold.
2. If it is lower than the threshold, we divide the sentences having a score higher than the threshold by commas and calculate the score of each part.

In the experiment, we set the threshold to 0.3.

2.1.4 Generation

The generation block produces an the essay by selecting some of the textbook sentences scored by the degree of similarity and labeled temporal relations in the following steps.

1. Select the highest-scored textbook sentences labeled "Match" while the essay fits within the character limit
2. Select the highest-scored textbook sentences labeled "Unknown" while the essay fits within the character limit
3. Select the highest-scored textbook sentences labeled "Mismatch" while the essay fits within the character limit

Finally, the selected sentences are sorted by their temporal information and output as the answer.

3. LONG ESSAY-TYPE QUESTIONS

The long essay-type questions usually specify several terms that must be used be in the answer. Our system for the long essay-type questions answers by using these terms because textbook sentences that include them are likely to be the

sentences to include in correct answer. The system is almost the same as the short essay-type question answering system. However, the score is modified a little. In equation (1), q is the set of the specified terms, and c is always 0 because we currently have no way to differentiate the importance of the terms.

4. FACTOID ANSWERING SYSTEM

Figure 3 shows the block diagram of our factoid QA system. In the category prediction block, the question is classified into the 21 types. Out of 21types, 20 types are defined in EVT and one is for those in none of them. We also categorized entries of world history glossary in advance by the category prediction block. In the search block, an answer is generated by searching entry name of the world history glossary based on the score described in Sec. 2.1. We thus describe only the category prediction in the next section.

4.1 Category prediction

This block classifies the question into the 21 categories. We semi-automatically extracted nouns from wikipedia pages of each world history term and made a dictionary for category prediction. For example, the page of Eratosthenes, categorized as "Person" in EVT, describes him as 「エラトステネス（Eratosthenes, 紀元前 275 年 - 紀元前 194 年）は、ヘレニズム時代のエジプトで活躍したギリシャ人の学者であり、アレクサンドリア図書館を含む研究機関であるムセイオンの館長を務めた。」. We extracted nouns 「学者」, 「館長」 from this page by using several regular expression patterns. We thus made a dictionary of nouns that indicate "Person", "Religion", "Location",etc., by collecting the keywords from the wikipedia pages for the terms categorized to those types in EVT. This block judges the categories by using this dictionary. For example, in the question 「神聖文字を解読したフランスのエジプト学者の名前を記しなさい.」, a noun 「学者」 indicating "Person" appears in it. This question is hence categorized as the "Person" type. We also categorized the world history glossary by using the same dictionary. This block outputs a category type and the next block searches for an glossary item on the basis of this output and the similarity score.

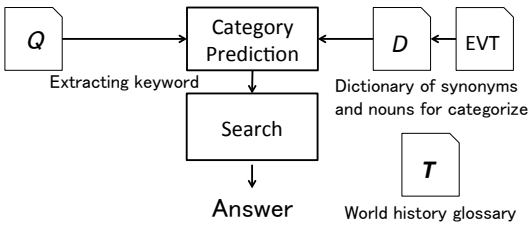

Figure 3: Factoid system block diagram

5. MULTIPLE-CHOICE QUESTION ANSWERING SYSTEM

5.1 Baseline System

Our system for answering multiple-choice question is based on "Surface-Similarity Based Textual Entailment Recognition for Japanese Text" [2] by Hattori et al. Figure 4 presents the outline of the system. The left side of Figure 4 is the system of Hattori et al. The right side of Figure 4 is the enhancement by us. First, the system of Hattori et al. calculates the degree of surface similarity between the choice and the paragraphs of textbooks. For each choice, it finds the paragraph with the highest degree of similarity to the choice. The degree of similarity is calculated by the scoring function in equation (1).

The system then extracts named entities and time information from the choice and the paragraph having the highest degree of similarity. It checks whether all the named entities in the choice are included in the textbook paragraph. When the time information of the choice corresponds with one of the time information in the textbook paragraph, we regard the time information of the choice matches the time information of the textbook paragraph. For example, suppose that a choice includes time information 「1946 年」and 「1950 年代」, and a paragraph of the textbooks includes time information such as 「1894 年」 and 「20 世紀」. This time information of the choice matches with the time information of the textbook paragraph because 「1946 年」 and 「1950 年代」are both included in 「20 世紀」.

When the degree of similarity exceeds a threshold, and the named entities and the time information of the choice matches those of the textbook paragraph, the choice is judges as being true.

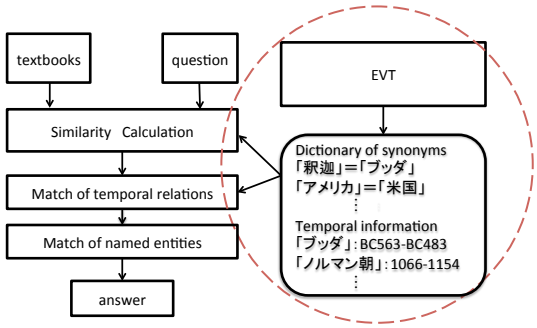

Figure 4: Overview of the answering system for the multiple-choice questions

5.2 Improvement

On the basis of an analysis of Hattori et al.'s system, we modified their scoring function as was shown in equation (1) and added the dictionaries described in Section 2 for rewriting nouns and extracting temporal information. We also used machine learning to integrate degree of surface similarity with the checking of time information and named entity (section 5.2.1). We also developed a module to answer the questions in which one chooses an correct temporal order among several historical events. We provide some details of the machine-learning based system below.

5.2.1 Scoring function base on Support Vector Machine

We developed a system based on SVM to synthesize the degree of surface similarity and the constraint checking about time information and named entities. Table 1 lists the features used in the system. We use a Support Vector Machine for Ranking [4] as the learning method.

Table 1: Feature values

Degree of surface similarity with the textbook (Yamakawa)
Degree of surface similarity with the textbook (Tokyo Syoseki)
Degree of surface similarity with the glossary (Yamakawa)
Result of the check of the time information in the textbook (Yamakawa)
Result of the check of the time information in the textbook (Tokyo Syoseki)
Result of the check of the named entities in the textbook (Yamakawa)

Table 2: Sundai exams

Phase 1	No	Score
Long essay	【1】	4/28
Short essay	1a	0/6
	1b	1/5
	2a	0/3
	2b	0/4
	3	0/4
Factoid	1-10	2/10

Phase 2	No	Score
Long essay	【1】	3/26
Short essay	1a	3/6
	1b	0/4
	2	1/6
	3a	0/4
	3b	0/4
Factoid	1-10	8/10

Phase 3	No	Score
Factoid	1-10	6/10

Our system calculates the feature vector for each choice using the textbooks and the glossary. The degree of similarity was calculated using equation (1). The system selects the choice with the highest score given by the SVM, i.e., the inner product of its feature vector and the weight vector.

6. EXPERIMENTS

6.1 Secondary exams

In this section, we describe only the result of Sundai mock test of university of Tokyo History exam in each run. Table 3 shows the results. In phase 1, we didn't use the dictionary of synonyms and the temporal relation labeling to answer essay-type questions, and didn't use category prediction to answer factoid questions. It hence seems that the total score of phase 2 is higher than that of phase 1 by these effects.

6.2 National Center Test

Table 3 shows the experimental results on National Center Tests and its mock test by Benesse. Benesse exam (using SVM) in phase 2 is the result by the system using SVM. All the other results are by the system using classification of nouns, synonym dictionary, and temporal relation labeling.

7. ANALYSIS AND DISCUSSION

We analyzed the results by the current short essay-type question answering system as well as the necessary knowledge and methods to improve it.

7.1 Mismatch of nouns by differences of abstraction levels

We found that in many cases appropriate sentences in the textbook are scored low because the description of the question is abstract while the textbook provides a concrete description on that topic. In such cases, the nouns of the question mismatch those of the textbook sentence. An examples of such a question, the model answer, and the most appropriate sentence in textbook are:

Table 3: Result

Phase 1	Score	Correctness
Center Test	38/100	15/41
Phase 2	**Score**	**Correctness**
Benesse exam	48/100	17/36
Benesse exam (using SVM)	41/100	15/36
Phase 3	**Score**	**Correctness**
Center Test	47/100	17/36

Question 明代の長江流域の農業・工業について，2 行以内で説明しなさい．(Sundai mock test 3-(b) in phase 2)

Model answer 下流域で綿織物など家内制手工業や綿花などの原料栽培が広がり，中流域が穀倉地帯となり「湖広熟すれば天下足る」と称された．

Textbook 長江下流域では綿織物や生糸に代表される家内制手工業がさかんになり，原料となる綿花や養蚕に必要な桑の栽培が普及した。

The above question includes abstract nouns such as 「農業」 and 「工業」. They correspond to specific nouns in the model answer such as 「綿織物」「家内制手工業」「綿花」「原料栽培」 and 「穀倉地帯」 similarly, the nouns 「農業」 and 「工業」do not appear in the textbook sentence, but specific nouns appear instead. Our system hence gave a low score to the appropriate textbook sentence. Such examples were particularly frequent in Tokyo University exams. This problem is presumably one of the reason for the lower score in Sundai Tokyo university mock exam than that of Keio University exam. For such cases, we need to utilize the knowledge that the nouns 「綿織物」,「家内制手工業」,「綿花」,「原料栽培」and「穀倉地帯」 are hyponym of the nouns 「農業」and 「工業」in the process of sentence extraction.

8. CONCLUSION

This paper described our system for world history essay and multiple-choice exams and the result in NTCIR-12 QALab. We discussed problems in answering short essay-type questions. We found that it is necessary to use knowledge tuned to domains (e.g., hypernym-hyponym relations among terms).

9. REFERENCES

[1] Noriko Arai and Takuya Matsuzaki. Can a robot get into the university of tokyo? *JSAI Journal*, Vol. 27, No. 5, pp. 463–469, 2012 in Japanese.

[2] Shohei Hattori and Satoshi Sato. Surface-similarity based textual entailment recognition for japanese text. *JSAI Journal*, Vol. 29, No. 4, pp. 416–426, 2014 in Japanese.

[3] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. Task Overview for NTCIR-12 QA Lab-2. In *Proc. NTCIR-12 Conference*, 2016.

[4] Thorsten Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 217–226. ACM, 2006.