SML Question-Answering System for World History Essay Exams at NTCIR-13 QALab-3

Yusuke Doi
Nagoya University
y_doi@nuee.nagoya-u.ac.jp

Takuma Takada
Nagoya University
takuma_t@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 question-answering system for world history short essay-type question at NTCIR-13 QALab-3 [1]. Our system consists of an extraction module and an compression module. In the extraction module, we identify the theme and the focus of a question, and extract several sentences from a glossary of world history that are appropriate for the theme and the focus. In the compression module, we compared three compression methods based on manually-designed compression rules, statistics from a corpus, and a hybrid of the both.

Team Name

SML

Subtasks

Term question, Essay question, End-to-End

Keywords

NTCIR-13, question answering, university entrance examination, world history, essay question, Theme, Focus

1. INTRODUCTION

We developed a system for solving the world history exams of the University of Tokyo. The questions in the exams are classified into three types: long essay, short essay, and factoid questions. We created a new question-answering system for the short essay-type questions. Figure 1 provides an example of the short essay questions and a model answer.

In our previous system [2], we extracted sentences from knowledge sources based on the surface similarity to the question sentence. The system could not deal with a question such as: "[1]Describe the league of cities established in Northern Italy, using no more than 30 English words," in which the theme of the question (i.e., Lombardia Alliance) is not mentioned explicitly. When the theme is not written in the question, it is difficult to extract appropriate sentences by using the surface similarity to the question. In addition, we did not consider the focus of the question. For instance,

ポリスの形成過程を、60字以内で説明しなさい。(Describe, in no more than 30 English words, the process by which the polis were formed.)

各地で有力貴族の指導下で、集落が連合し、アクロポリスを中心として人々が集住する形でポリスを形成した。
(Under the leadership of powerful nobles, various settlements formed coalitions, and people lived together around the Acropolis, forming poleis.)

Figure 1: A short essay-type question and a model answer[2]

the following question asks about the reason but not the result of a historical fact: "Explain the reason that Christians were persecuted the Roman Emperor within 60 characters." We hence need to extract a sentence describing the reason but not the result of the persecution of Christians in Roman Empire, which is not easy only with the surface similarity.

Another problem in the previous system was that the answer was made from the extracted sentences simply by dividing them at the punctuation marks and selecting the segments most similar to the question. However, to achieve a higher score, it is necessary to make a short and good answer that includes important information scattered across the extracted sentences by deleting unnecessary words.

The new system consists of two modules: the extraction module (§2.1) and the compression module (§2.2). In the extraction module, we identify the theme and the focus of the question. Next, an item is selected from a glossary of world history, and the sentences in the item which match the focus of the question are extracted. We compared three methods in the compression module: a rule-based method, a query-oriented summarization method, and a hybrid method. The rule-based method compresses extracted sentences according to several compression rules. The query-oriented summarization method compresses the extracted sentences based on the Query Snowball (QSB) [3]. The hybrid method compresses the extracted sentences using both the rules and the QSB score.

The rest of this paper is organized as follows. Section 2 describes the question-answering system. Section 3 describes the evaluation results. Section 4 provides an analysis and discuss the evaluation results.

[1]<K792W10-9> The University of Tokyo, 2009

2. SYSTEM

In this section, we first describe how we identify the theme and the focus of a question to extract appropriate sentences. We then explain the three compression methods.

2.1 Extraction Module

First, we analyze the question sentence using KNP[3], and extract the nouns which depend on specific phrases such as "述べよ (describe)" and "説明せよ (explain)." We identify the categories of the theme and the focus of the question from those nouns. For example, when the question is "Describe the league of cities established in Northern Italy, using no more than 30 English words," the category of its theme is "Organization" and the focus is "Activity."

In order to identify the category of the theme, we made a nouns dictionary that consists of 980 nouns classified into 23 categories. The categories are those defined in the World History Event Ontology (EVT)[4] [4]. The nouns were extracted from the first sentences of the Wikipedia pages of the items defined in the EVT. For example, "Eratosthenes" is classified into the "Person" category in the EVT. Nouns such as "mathematician" and "poet" are extracted from the first sentence of Wikipedia page of the "Eratosthenes":

> Eratosthenes of Cyrene was a Greek mathematician, geographer, poet, astronomer, and music theorist.

These nouns are used as the keywords that indicate the theme of the question is under the "Person" category.

We manually classified the items in the glossary into 23 categories. We extract the items which are in the same category as the theme. We also extract time and location expressions from the question and exclude the items that include a time or location expression that does not match those in the question. For each of the remaining items, we calculate the surface similarity between its explanation sentences and the question sentences. We select the item having the highest similarity as the theme. If more than one item have the same, highest similarity, all these items are identified as the theme. If the nouns extracted from the question has an item in the glossary, we simply regard the noun as the theme.

We identify the focus by using the nouns extracted from the question. We manually made a list of nouns expressing a focus in advance (Table 1). If the list includes the extracted noun, the focus associated to it is identified as the focus of the question. If the list does not include any of the extracted nouns, "Content" and "Activity" are identified as the focus of the question.

Each explanation sentences in the glossary was given one or more focus labels by pattern matching in advance. For example, if a sentence include any of "特徴 (character)", "特色 (character)", "側面 (aspect)", and "性格 (nature)," it is given a focus label of "特徴 (Character)." We extract the explanation sentences of the theme which have the same label as the focus. If there are no such sentences, we extracts all the explanation sentences.

2.2 Compression Module

Table 1: Nouns expressing a focus

focus	noun
内容 (Content)	内容 (content)
理由 (Reason)	理由 (reason), 誘因 (incentive)
結果 (Result)	結果 (result)
確立 (Establishment)	確立 (establishment), 経緯 (process)
過程 (Process)	過程 (process), 経緯 (process)
変化 (Change)	変化 (change), 変遷 (transition)
情勢 (Situation)	情勢 (situation), 動向 (trend)
特徴 (Character)	特徴 (character), 特質 (character)
活動 (Activity)	活動 (activity), 行う (perform)
役割 (Role)	役割 (role)

In the compression module, we make the answer by compressing the sentences extracted by the extraction module. We compared three compression methods.

2.2.1 *Rule-Based Method*

The algorithm of rule-based method is as follows:

1) Sort extracted sentences by the overlap of nouns between the sentence and the question.

2) Create a list of answer candidates from the sorted extracted sentences. When the sentences sorted in descending order are $s_1, s_2, \cdots, s_n$, the answer candidates are $s_1 + s_2, s_1, s_2, \cdots, s_n$. $s_1 + s_2$ is made by combining s_1 and s_2.

3) Replace words with shorter synonyms.

4) Compress the first answer candidate by using the compression rules described below. If the number of characters in the compressed sentence is less than the character limit, output it as the answer.

5) If the compressed sentences is longer than the character limit, compress the next answer candidates by using the compression rules.

The compression rules are:

a) Delete parenthesized phrases

b) Delete conjunctions at the beginning of the sentences.

c) Delete adverbs.

d) Delete time expressions.

e) Delete adnominal phrases one by one from the beginning of the sentence.

We repeat the algorithm until the system outputs a sentence for the answer. If the number of characters of the last compressed sentence is more than the character limit, the output is blank.

2.2.2 *Query-Oriented Summarization Method*

The second method is based on the subtree extractive summarization via submodular maximization proposed by Morita et al. [5]. First, we analyze the extracted sentences by using KNP. We repeatedly add valid subtrees to the answer while the number of characters in the answer is less than the characters limit. The valid subtrees are the subtrees of a dependency tree including the root of the sentence. At each

[2]<K792W10-2> The University of Tokyo, 2009
[3]http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[4]http://researchmap.jp/zoeai/event-ontology-EVT/

step, a valid subtree is selected and added to the answer so that it maximizes the gain in an objective function.

The objective function is the score of a provisional summary defined as follows:

$$f(S) = \sum_{w \in \text{words}(S)} \left\{ \text{qsb}(w) \sum_{i=0}^{\text{count}_S(w)-1} d^i \right\} + \gamma \ \text{reward}(S) \quad (1)$$

$$\text{reward}(S) = c(S) - |S| \quad (2)$$

where d is the damping rate, $\text{count}_S(w)$ is the number of sentences containing word w in summary S, $\text{qsb}(w)$ is the query relevance score of w, γ is a parameter that adjusts the rate of the compression, $c(S)$ is the number of characters in S, and $|S|$ is the number of sentence in S. The first term of Eq.(1) expresses the sum of the query relevance scores of the words in S, and the second term expresses a reward for readability. We gradually damp the query relevance scores of the words that already appear in the answer. The reward leads to a natural summary being generated with fewer sentences and penalizes too short sentences.

2.2.3 *Hybrid Method*

We designed a hybrid method that combines the rule-based method and the QSB score. In the rule-based method, we make the list of answer candidates from the extracted sentences sorted in the descending order of the overlap of nouns between the sentence and the question. In the hybrid method, we try compressing each of the extracted sentences and the concatenation of each pair of them. That is, when the extracted sentences are $s_1, s_2, \ldots, s_n$, the answer candidates are:

$$\{s_1, s_2, \ldots, s_n\} \cup \bigcup_{i<j} \{s_i + s_j\}$$

The list of answer candidates is made by sorting them in the descending order of their QSB scores. Thereafter, they are compressed in turn in the same manner as the rule-based method. Because answer candidates of this method are more than those of rule-based method, we can output the answer among more choices of answer candidates.

3. RESULT OF THE TEST RUN

We participated in the phase 2 of the test-run. We submitted the answers to the short-essay questions and the factoid questions. The answer to the factoid questions were produced by our previous system [2]. For the short-essay questions, we submitted two runs. In the run 1, we used the hybrid compression method. In the run 2, we used a method based on the query-oriented summarization, where we deleted conjunctions at beginning of the sentences and the time expressions before the compression. We manually evaluated the answers by using the answer nuggets provided by the task organizers. Table 2 shows the result. The number of the nuggets of covered by the two runs was the same. The ROUGE scores against reference answer were also very close between the two runs.

Table 2: Short essay result of phase2

runs	nuggets	ROUGE-1	ROUGE-2	ROUGE-3
run1	7/80	0.313	0.088	0.038
run2	7/80	0.312	0.091	0.039

4. ANALYSIS

In 18 out of 22 questions, none of the extracted sentences were appropriate for the question. On the majority of them (8 questions), the question sentences explicitly mention to a theme but the system failed to extract it. For instance, the system failed to identify the correct theme "儒学 (Confucianism)" on the following question:

> **Question:**[5] それまで複数の有力な思想の一つにすぎなかった儒学が，他の思想とは異なる特別な地位を与えられたのは，前漢半ばであった。そのきっかけとなった出来事について 60 字以内で説明しなさい。
> (During the middle of the Former Han era Confucianism, which up until that point had been merely one of several valid schools of thought, was given a special position of prominence, separate from other schools of thought. Explain, in 30 English words or less, what event led to this.)

"儒学 (Confucianism)" was not extracted as the theme because it does not depend on the key phrase "説明せよ (explain)."

On the following question, the system also failed to identified the correct theme "ツンフト闘争 (Zunftkämpfe)".

> **Question:**[6] 西ヨーロッパでは中世都市が発展すると，おもに手工業生産者からなるツンフトとよばれる組織が形成され，彼らが主体となるツンフト闘争が各地で起こった。この闘争は誰に対する何を求めた闘争だったか。30 字以内で記述しなさい。
> (In Western Europe, the development of medieval cities led to the formation of organizations, called Zunft, consisting primarily of handicraftsmen. The Zunft which they led engaged in battles in various regions. In 15 English words or less, explain who these battles were against, and what they demanded.)

If a key word of world history is in the question, we should identify the key word as the theme.

In 4 out of 22 questions, we could extract sentences that include at least one of the nuggets. We analyzed to compare the hybrid compression method with the query-oriented summarization method. There were differences in the nuggets covered by the two methods on 2 questions. These questions are shown bellow.

> **Question:**[7] 日本は下線部①の連合組織に参加し，後に脱退した。脱退の経緯を 60 字以内で記せ。
> (Japan participated in the federation in underlined section (1), but then left it. Explain, in 30 English words or less, what led to leaving the federation.)
> **Nugget:** 日本はリットン調査団の報告を不服とした。(Japan challenged the Lytton Commission's report.)
> **Extracted sentence:** 国際連盟脱退は1933 年 3 月、日本は、連盟総会でリットン調査団報告に基づく満州撤兵などの勧告案が、42 対 1 で採択された

[5] <L792W10-2> The University of Tokyo, 2010
[6] <P792W10-11> The University of Tokyo, 2014
[7] <B792W10-2> The University of Tokyo, 2001

のを不満として脱退した。
Run1: 国際連盟脱退は日本は、連盟総会で基づく満州撤兵などの勧告案が、42 対 1 で採択されたのを不満として脱退した。
Run2: 国際連盟脱退は日本は、リットン調査団報告に基づく満州撤兵などの勧告案が、42 対 1 で採択されたのを不満として脱退した。

Each extracted sentence are was prepended with "item name + は" at the beginning because almost all of the sentences do not have a subject. The nugget is included only in the answer of Run 2 because the compression rule (d) deleted the adnominal phrase that is necessary for the answer.

Question:[8] 西アジアのアラビア半島では，ワッハーブ派が勢力を拡大した。この運動について 90 字以内で説明しなさい。
(In the Arabian Peninsula, in western Asia, the Wahhabists grew in power. Describe, in 45 English words or less, this movement.)
Nugget: ワッハーブ派は、サウード家と結んだ。
(Wahabbists joined with the House of Saud.)
Run1: ワッハーブ派はサウード家と結んで勢力を拡大した。ワッハーブ派はムハンマド時代のイスラム教への回帰をとなえ、神秘主義やシーア派を厳しく批判し、シャリーアの厳密な適用を図ろうとした。
Run2: ワッハーブ派は始められたイスラーム教の改革派。ワッハーブ派はムハンマド時代のイスラーム教への回帰をとなえ、神秘主義やシーア派を厳しく批判し、シャリーアの厳密な適用を図ろうとした。

The nugget is included only in the answer of Run 1, because the QSB score of "サウード家 (house of Saud)" was zero and hence deleted. On these question, beause the average number of the extracted sentences was about 1.8, they do not have to be compressed and there was little difference between the answer of the two run. We cannot conclude which compression method is better.

5. CONCLUSION

This paper described the question-answering system for the world history short essay-type questions and the result of phase 2. There was no significant difference between the performances of a rule-based and statistics-based compression methods. The analysis of the test-run results revealed a need for a more accurate detection of the theme of the question.

6. REFERENCES

[1] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kondo. Overview of the NTCIR-13 QA Lab-3 Task. In *13th NTCIR Conference*, 2017.

[2] Takuma Takada, Takuya Imagawa, Takuya Matsuzaki, and Satoshi Sato. Sml question-answering system for world history essay and multiple-choice exams at ntcir-12 qalab. In *12th NTCIR Conference*, 2016.

[3] Hajime Morita, Tetsuya Sakai, and Manabu Okumura. Query snowball: A co-occurrence-based approach to multi-document summarization for question answering. In *Proceedings of ACL-HLT'11 - Volume 2*, pp. 223–229, 2011.

[4] Ai Kawazoe, Yusuke Miyao, Takuya Matsuzaki, Hikaru Yokono, and Noriko Arai. *World History Ontology for Reasoning Truth/Falsehood of Sentences: Event Classification to Fill in the Gaps Between Knowledge Resources and Natural Language Texts*. 2014.

[5] Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Subtree extractive summarization via submodular maximization. In *Proceedings of ACL'13*, pp. 1023–1032, 2013.

[8] <L92W10-6> The University of Tokyo, 2010