Automatic Evaluation of World History Essay Using Chronological and Geographical Measures

Kotaro Sakamoto
Yokohama National University
National Institute of Informatics
sakamoto@forest.eis.ynu.ac.jp

Hideyuki Shibuki
Yokohama National University
shib@forest.eis.ynu.ac.jp

Madoka Ishioroshi
National Institute of Informatics
ishioroshi@nii.ac.jp

Akira Fujita
Yokohama National University
fujita@ynu.ac.jp

Yoshinobu Kano
Shizuoka University
kano@inf.shizuoka.ac.jp

Teruko Mitamura
Carnegie Mellon University
teruko@cs.cmu.edu

Tatsunori Mori
Yokohama National University
mori@forest.eis.ynu.ac.jp

Noriko Kando
National Institute of Informatics
SOKENDAI
kando@nii.ac.jp

ABSTRACT

We propose a method for measuring chronological and geographical consistency of the world history essays in Japanese university entrance exams. The experimental result shows a weak positive correlation between the scores measured by the proposed method and the scores estimated by a human expert in world history.

KEYWORDS

essay QA, automated evaluation, chronological and geographical measures, world history, university entrance exams

1 INTRODUCTION

Research on real-world complex question-answering (QA) has flourished in recent years [1]. In the QA Lab tasks [10] at the NTCIR workshop,[1] the current problems and solutions in QA technologies have been investigated using the world history questions in Japanese university entrance exams and their English translation. Japanese university entrance exams include various types of questions such as multiple-choice, fill-in-the-blank, true-or-false and essay questions. Above all, essay QA is the most challenging, and still has many open problems, such as the evaluation of essays that QA systems generated. Although there is a way of evaluation by human experts in world history, it takes considerable time and cost. In the case of the QA Lab, evaluation of 46 essays by an expert who teaches world history took around a month and about 500,000 yen (4,500 USD). Therefore, a new method is required.

Because essay generation is regarded as a kind of query-biased summarization, the measures for evaluating summaries using gold-standard data can be applied to essay evaluation. In the QA Lab, the ROUGE family [5] and the Pyramid method [7, 9] are used for grading essays besides a human expert's evaluation. A positive correlation between these grades and those provided by humans was between moderate and weak, and the ranking order by the measures was not always concordant with the ranking order given by the human marks. Therefore, we investigated more appropriate measures for evaluating world history essays in Japanese university entrance exams.

For evaluating summaries, the linguistic well-formedness and the relative responsiveness were used in the DUC workshops.[2] The content, readability/fluency, and the overall responsiveness were used at the Guided Summarization tasks[3] in the TAC workshops. These measures are important for evaluating world history essays in university entrance exams. However, the linguistic well-formedness and readability/fluency were scored arbitrarily by human assessors, while the content was methodologically scored by the ROUGE family and the Pyramid method, among others. We would like to methodologically give other scores based on merits other than the content. For evaluating world history essays, chronological and geographical consistency is important as a kind of semantic consistency. However, how to evaluate these is not obvious. In this paper, we propose a method for measuring chronological and geographical consistency of world history essays, and examined the method using essays submitted to the QA Lab.

The main contributions of this paper are as follows: (i) to clarify the features of well-formed world history essays in terms of the chronological information and the geographical information, (ii) to introduce a new scoring method based on the features to evaluate the well-formedness of world history essays.

2 RELATED WORK

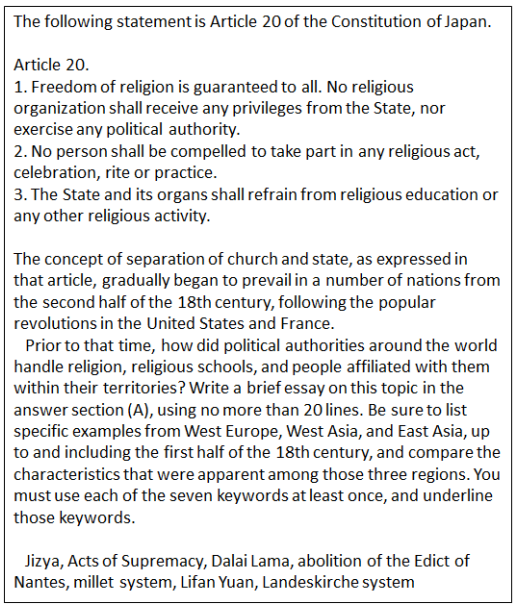
The linguistic well-formedness in the DUC workshop and the readability/fluency in the TAC Guided Summarization tasks were evaluated in terms of grammaticality, non-redundancy, referential clarity, focus, and 'structure and coherence'. Our measures are relative to the focus and 'structure and coherence'. Although Barzilay et al. [2] and Okazaki et al. [8] researched the chronological ordering, they did not take account of geographical information. Buscaldi et al. [4] found that geography is related to semantic similarity, but they only aimed to measure semantic equivalence between two

[1] http://research.nii.ac.jp/ntcir/index-en.html

[2] http://duc.nist.gov/duc2007/tasks.html

[3] http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html

EVIA 2017, co-located with NTCIR-13, 5 December 2017, Tokyo, Japan.


The following statement is Article 20 of the Constitution of Japan.

Article 20.
1. Freedom of religion is guaranteed to all. No religious organization shall receive any privileges from the State, nor exercise any political authority.
2. No person shall be compelled to take part in any religious act, celebration, rite or practice.
3. The State and its organs shall refrain from religious education or any other religious activity.

The concept of separation of church and state, as expressed in that article, gradually began to prevail in a number of nations from the second half of the 18th century, following the popular revolutions in the United States and France.
Prior to that time, how did political authorities around the world handle religion, religious schools, and people affiliated with them within their territories? Write a brief essay on this topic in the answer section (A), using no more than 20 lines. Be sure to list specific examples from West Europe, West Asia, and East Asia, up to and including the first half of the 18th century, and compare the characteristics that were apparent among those three regions. You must use each of the seven keywords at least once, and underline those keywords.

Jizya, Acts of Supremacy, Dalai Lama, abolition of the Edict of Nantes, millet system, Lifan Yuan, Landeskirche system

Figure 1: An example of essay question of world history

text snippets. Because Madanani et al. [6] only researched sentence ordering, the research only applied to the context of a short, domain-independent summarization. Bauer and Teufe [3] proposed the extended Pyramid method for timeline summarization, but they did not focus on the well-formedness. Although Wagner et al. [11] researched the well-formedness, they focused only on grammatical errors. Therefore, there is no research on a methodology for measuring the focus and the structure and coherence of world history essays in terms of the chronological and geographical information.

3 ESSAY QUESTION OF WORLD HISTORY

Figure 1 shows an example of an essay question for world history, which is an English translation from the original Japanese version. The question contains additional text besides the main essay topic. The first paragraph gives background information, and the texts below the essay topic are the constraints for writing the essay. The constraints include a length limitation of "no more than 20 lines," a geographical condition of "West Europe, West Asia and East Asia," a chronological condition of "up to and including the first half of the 18th century," the keywords that must be used in the essay, and other associated conditions. The chronological condition and the geographical condition prove the importance of chronological and geographical consistency.

4 WELL-FORMED WORLD HISTORY ESSAY

4.1 Structure

In general, a world history essay is a sequential description of historical events (HEs). A HE has both chronological information and geographical information. Let us consider how this is written. While the chronological information can be easily put in a linear order from the past to the future, the geographical information is not easy to be determinately put in a linear order because of the spatial extent. Based on the study of several model answer essays from past university entrance exam collections, the general structure of the essays follows one of two approaches: (a) disregarding geographical information, all HEs are described in chronological order, and (b) grouping HEs by the geographical information. In both, information is described in chronological order. If the former is regarded to be grouped by geographical information from "the whole world," there is no difference between the two manners; that is, both are descriptions in chronological order for HEs in a particular area. We defined a sequence of HEs with the same geographical information as a geographical section (GS). GSs could be nested hierarchically. For example, a GS of Europe may contain GSs such as England, France, and Germany, and the GS of England may contain GSs such as London, Birmingham and Manchester.

From the above, we built the following hypotheses for the structure of world history essay.

(H1) An essay is a GS.
(H2) A GS can consist of more than one sub-GSs that is in the parent GS.
(H3) HEs in a GS are put in chronological order.

GSs in a hierarchical structure are classified into terminal and non-terminal sections. A terminal section means an HE sequence without hierarchical structure, and likewise a non-terminal section can be divided into several GSs. We defined a non-terminal section corresponding to the essay as the root section. A GS s is defined as a paired HE sequence $E = (e_1, e_2, \cdots, e_m)$ and GS sequence $SS = (s_1, s_2, \cdots, s_n)$. If SS is an empty tuple, then the GS is a terminal section. HEs in a sub-GS are shared with the superordinate GS, and E of non-terminal sections are not empty. For a question, the chronological condition CC is defined as a pair of the beginning time bt and the ending time et, and the geographical condition GC is defined as a geographical entities set $\{g_1, g_2, \cdots, g_k\}$.

4.2 Uniformity

Let us consider the uniformity of GSs in a GS. If GSs of the East Midlands, Paris and German are placed on the same level in a GS of Europe, they are incongruous even though they are all parts of Europe. This is because they are in different levels of a geographical category, such as country, region, and city. Therefore, well-formed essay require the uniformity of geographical category level. In addition, if England is described with hundreds of words while France and Germany are respectively described with a dozen words, there is incongruity even though they are in the same geographical category level. This is because their quantities of description are imbalanced. Therefore, well-formed essay seems to require the uniformity of quantity.

We built the following hypotheses for the uniformity of GSs.

(H4) GSs placed on the same level in a GS are in the same level of geographical category.
(H5) GSs placed on the same level in a GS are described in the same quantity.

Although several functions implementing the hypotheses were come up with, we took simple functions by way of experiment. The geographical uniformity $sc_{GU}()$ and the quantity uniformity $sc_{QU}()$ are calculated by the following functions:

$$sc_{GU}(SS) = 1 - \frac{sd_{GU}(S)}{am_{GU}(SS)} \quad (1)$$

$$sd_{GU}(SS) = \sqrt{\frac{1}{|SS|}\sum_{i=1}^{|SS|}(depth(s_i) - am_{GU}(SS))^2} \quad (2)$$

$$am_{GU}(SS) = \frac{1}{|SS|}\sum_{i=1}^{|SS|} depth(s_i) \quad (3)$$

$$sc_{QU}(SS) = \frac{-\sum_{i=1}^{|SS|} p(s_i, SS)\log_2 p(s_i, SS)}{\log_2 |SS|} \quad (4)$$

$$p(s, SS) = \frac{length(s)}{\sum_{i=1}^{|SS|} length(s_i)} \quad (5)$$

where $depth(s)$ is a function to return the distance between the thesaurus root node and the node corresponding to the range of s. $length(s)$ is a function to return the number of characters described in s. We designed the scoring functions to be normalized into the range [0, 1].

4.3 Ordering

Let us consider the ordering of HEs in a GS. HEs in well-formed essays are generally described in chronological order. Note that the occurrence order of HEs does not always correspond with the descriptive order of an essay. Since the chronological information of an HE has a beginning and ending in a range, the occurrence order relation between HEs is either non-overlapping, partially overlapping or inclusive.In all relations, the beginning of the HE e_1 precedes the beginning of the HE e_2. However, in the inclusion relation, e_1 may be described after e_2 such as "The Treaty of Nanking ended the First Opium War." Therefore, we assume that the describing order of HEs in the inclusion relation is free to the chronological order. Next, let us consider the ordering of GSs in a GS. The describing order of GSs is free relative to the chronological order. However, for example, the describing order of Athens, Rome, Cairo, Baghdad, Beijing and Shanghai seems to be better than the order of Athens, Baghdad, Beijing, Cairo, Rome and Shanghai. This is because GSs relating to each other are placed closely. We assume that the relativity is approximated by the geographical distance.

We built the following hypotheses for the ordering in a GS.

(H6) As an exception to the hypotheses (H3), an HE can be described both before and after another HE if they are in the inclusion relation.

(H7) GSs in a GS are described in the order of short geographical distance.

The hypothesis (H6) is the complement of the hypothesis (H3).

The chronological ordering $sc_{CO}()$ and the geographical ordering $sc_{GO}()$ are calculated by the following functions:

$$sc_{CO}(E) = \frac{K - L}{K + L} \quad (6)$$

$$sc_{GO}(E) = \frac{1}{geochange(E) + 1} \quad (7)$$

$$geochange(E) = \frac{1}{|E| - 1}\sum_{i=1}^{|E|-1} distance(range(e_i), range(e_{i+1})) \quad (8)$$

where K is the number of concordant pairs of HEs in E, and L is the number of discordant pairs. $range(e)$ is a function to return a thesaurus node that is the nearest common node subsuming all geographical entities included in the HE e, and $distance(n_i, n_j)$ is a function to return the shortest distance between the thesaurus nodes n_i and n_j.

4.4 Cooperability

Let us consider the cooperability of a world history essay to question constraints in terms of the chronological and the geographical information. As described in Section 3, world history essay questions give chronological and geographical conditions such as "up to and including the first half of the 18th century" and "West Europe, West Asia and East Asia." In this case, if an essay describes only the ancient histories of West Europe, West Asia and East Asia, the essay satisfies the conditions logically. However, it does not reflect the question intention. Cooperative essay should describe at least one HE of the 18th century. The geographical information is also similar. For example, an essay describing only "West Europe and West Asia" violates the maxim of quantity, and the cooperative essay should describe at least one HE for each area of the geographical condition. We assume that the chronological cooperability is observed in all GSs while the geographical cooperability is observed in only a GS corresponding to the essay. For a GS, we defined a period from the beginning of the earliest HE to the end of the latest one as a period of the GS. The smallest geographical range, including where the HEs in a GS occurred, was defined as the range of the GS. We assume that the observance of the maxim of quantity is approximated to the coverage of the period and the range of GSs.

We built the following hypotheses for the cooperability on the chronological and the geographical conditions in questions.

(H8) A period of a GS covers the period of the chronological condition as justly as possible.

(H9) A range of a GS corresponding to the essay covers the range of the geographical condition as justly as possible.

The chronological cooperability $sc_{CC}()$ and the geographical cooperability $sc_{GC}()$ are calculated by the following functions:

$$sc_{CC}(E, CC) = \frac{overlap(period(E), CC)}{extend(period(E), CC)} \quad (9)$$

$$sc_{GC}(E, GC) = \frac{2P(E, GC)R(E, GC)}{P(E, GC) + R(E, GC)} \quad (10)$$

$$P(E, GC) = \frac{subsumed(geoentities(E), GC)}{|geoentities(E)|} \quad (11)$$

$$R(E, GC) = \frac{subsuming(geoentities(E), GC)}{|GC|} \quad (12)$$

where $period(E)$ is a function to return a pair of the earliest time and the latest time in E, $overlap(P_1, P_2)$ is a function to return the length of the overlap period between P_1 and P_2, and $extend(P_1, P_2)$ is a function to return the length of the period between the earliest time and the latest time among P_1 and P_2. $geoentities(E)$ is a function that returns a set of geographical entities included in E, $subsumed(G_1, G_2)$ is a function that returns the number of geographical entities of G_1 subsumed by geographical entities of G_2, and $subsuming(G_1, G_2)$ is a function that returns the number of geographical entities of G_2 subsuming geographical entities of G_1.

5 PROPOSED METHOD

Figure 2 shows the outline of the proposed method. First, the input essay is segmented into HEs by punctuation marks. A HE is represented by a set of named entities extracted from the segment. Some

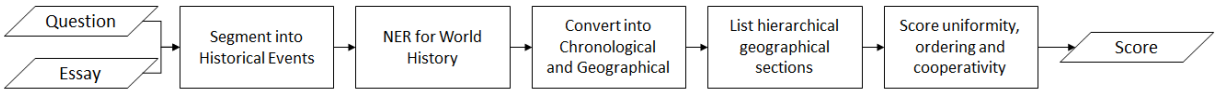


Figure 2: The outline of the proposed method

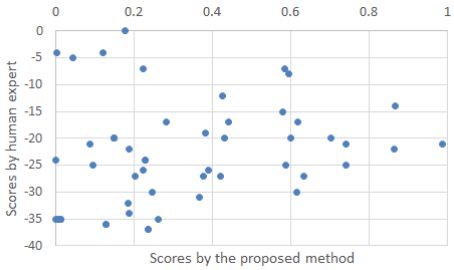


Figure 3: The scatter plot between the scores by the method and the scores by a human expert

named entities evoke the chronological and/or the geographical information. Because exam cram books cover such information, we constructed a database of world history terms based on the world history glossary published by Yamakawa Shuppan-sha.[4] Using the database, the named entities are converted into chronological and geographical information. Using both chronological and geographical information sets, the period and the range of the segment are respectively determined in the same way as that of the GS described in 4.4. They are regarded as the chronological and geographical information of the HE. Then, all hierarchical structures of GSs that can be gotten from the essay are listed. After scoring the HEs for each hierarchical structure, the maximum score is selected as the final score for the essay in order to select the most plausible hierarchical structure.

Based on the hypotheses described in Section 4, the score sc for a GS to a question is recursively calculated by the following functions.

$$sc(E, SS, CC, GC) = \begin{cases} sc_T(E, CC) \\ \quad \text{if it is a terminal section} \\ sc_N(E, SS, CC) sc_{GC}(E, GC) \\ \quad \text{if it is the root section} \\ sc_N(E, SS, CC) \\ \quad \text{otherwise} \end{cases} \quad (13)$$

$$sc_T(E, CC) = sc_{CO}(E) sc_{GO}(E) sc_{CC}(E, CC) \quad (14)$$

$$sc_N(E, SS, CC) = \frac{1}{|SS|} sc_{GU}(SS) sc_{QU}(SS) \sum_{i=1}^{|SS|} sc(events(s_i), sections(s_i), CC, GC) \quad (15)$$

where $events(s)$ and $sections(s)$ are respectively functions to return an HE sequence and a GS sequence included in a GS s.

6 EXPERIMENTAL RESULT

Using essays submitted to the QA Lab-2 and the QA Lab-3, we compared the scores measured by the proposed method and the scores evaluated by human expert. Although the number of the essays is 55, they are annotated with the marks granted and taken away besides the total score by a human expert. Basically the marks awarded take account of the correctness of the content, and the marks lost account for the ill-formedness. With this, we compared the scores to the method behind subtracting marks. Note that the lost marks are caused by not only chronological and geographical inconsistencies. Figure 3 shows the scatter plot between the scores by our method and the subtracted marks. The correlation coefficient was 0.21, which indicated a weak positive correlation. Taking into account that the marks subtracted include other causes than the chronological and geographical problems, the value seems to be fairly good.

[4]http://www.yamakawa.co.jp/ (in Japanese)

7 CONCLUSION

For world history essays in Japanese university entrance exams, we proposed a method for measuring the uniformity, ordering and cooperability in terms of the chronological and the geographical information. The features of well-formedness are found by observing several model answer essays. From the experimental result, we found a weak positive correlation between the scores measured by our method and the scores estimated by a human expert. We will investigate more appropriate functions in the future.

REFERENCES

[1] Eugene Agichtein, David Carmel, Donna Harman, Dan Pelleg, and Yuval Pinter. 2015. Overview of the TREC 2015 LiveQA Track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference.*

[2] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research* 17, 1 (2002), 35–55.

[3] Sandro Bauer and Simone Teufe. 2015. Improving Chronological Sentence Ordering by Precedence Relation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Vol. 2. 834–839.

[4] Davide Buscaldi, Jorge J. Garcia Flores, Joseph Le Roux, and Nadi Tomeh. 2014. LIPN: Introducing a new Geographical Context Similarity Measure and a Statistical Similarity Measure Based on the Bhattacharyya Coefficient. In *Proceedings of the 8th International Workshop on Semantic Evaluation.* 400–405.

[5] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out.* 74–81.

[6] Nitin Madnani, Rebecca Passonneau, Necip Fazil Ayan, John M. Conroy, Bonnie J. Dorr, Judith L. Klavans, Dianne P. O'Leary, and Judith D. Schlesinger. 2007. Measuring Variability in Sentence Ordering for News Summarization. In *Proceedings of the Eleventh European Workshop on Natural Language Generation.* 81–88.

[7] Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* 145–152.

[8] Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. 2004. Improving Chronological Sentence Ordering by Precedence Relation. In *Proceedings of the 20th International Conference on Computational Linguistics.* 81–88.

[9] Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated Pyramid Scoring of Summaries using Distributional Semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* 143–147.

[10] Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. 2017. Overview of the NTCIR-13 QA Lab-3 Task. In *Proceedings of The NTCIR-13 Conference.*

[11] Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2007. A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* 112–121.