DGLab Question Answering System and Automatic Evaluation Method at NTCIR-13 QA Lab-3 for University Entrance Exam on World History Essay

Mike Tian-Jian Jiang
DG Lab, Digital Garage, Inc.
Tokyo, Japan
miket@dglab.com

ABSTRACT

The paper describes DGLab question answering system and automatic evaluation method at NTCIR-13 QA Lab-3 for Japanese university entrance exam on world history essay. Submissions of subtasks include extraction, summarization, and evaluation method, in Phase-2 and Research Run for both Japanese and English. The proposed system follows the organizer-recommended pipeline, consisting of an extraction module and a summarization one. The former comprises a condition extraction routine, a full-text search engine, and a sentence selection heuristics. The latter one utilizes well-established summarization algorithms that extract sentences based one their importance. For evaluation method, an independent system measures semantic similarity between submitted summaries and gold standard essays. While particularly on summarization subtask, the submissions have gained competitive performance among participants, in terms of total scores graded by human experts. The quality of the machine-generated essay, however, is far from ideal to pass the exam. A brief discussion is therefore presented to diagnose symptoms and their probable causes.

Categories and Subject Descriptions

• **H.3.4 [INFORMATION STORAGE AND RETRIEVAL]:** Systems and Software - Performance evaluation (efficiency and effectiveness), Question-answering (fact retrieval) systems.

Keywords

DGLab; NTCIR-13; QA Lab-3; Question Answering; Essay Question; University Entrance Exam; World History; Extractive Summarization; Word Mover's Distance

Team Name

DGLab

Subtasks

Task for Essay Questions: Phase-2 (Extraction, Summarization, Evaluation Method) / (Japanese, English) subtasks; Research Run (End-to-end) / (Japanese, English) subtasks

1. INTRODUCTION

NTCIR (NII Testbeds and Community for Information-access Research) conference is composed of several shared tasks that dedicate to Information Access (IA) technologies such as information retrieval and extraction, question answering (QA), text summarization, etc. It is one of the major academic events in Asia-Pacific region, resembling to its counterparts CLEF, FIRE, TREC in Europe, America, and South-Asia, respectively.

Despite QA is usually considered as an advanced form of information retrieval, it is not yet as popular as search engines in terms of industrial and commercial usages. In attempts to tackle real-world problems, QA Lab[6–8], one of the NTCIR tasks, investigates complex QA technologies and appropriate evaluation methodologies by utilizing Japanese university entrance exams on world history as a touchstone, with a joint effort of participants. Japanese university entrance exam comprises two stages: The National Center Test of multiple choice questions and the college-specific Second-stage Test of both terms as objective assessment and essays as subjective one.

In the 13th NTCIR (NTCIR-13), the 3rd QA Lab (QA Lab-3)[7], like the preceding two[6, 8], consists of three types of questions, namely multiple-choice, term (usually as named-entity or terminology), and essay. While multiple-choice and term questions roughly fall into the category of factoid QA, which has been well studied for several decades, non-factoid and subjective QA such as essay remains relatively immature. One of the obstacles to rapid research and development on essay QA system is the labor-intensiveness of human evaluation due to its subjective nature. For example, according to QA Lab organizers, a world history teacher has spent about a month and billed approximately 500,000 yen (4,500 USD) to evaluate 46 essays[5]. Therefore, QA Lab-3 migrates to three independent subtasks, that resemble the lesson learned from DUC and TAC workshops, as follows: (passage) extraction, (query-biased/guided) summarization, and evaluation method, to provide incentives for participants to work individually on less complicated yet more focused goals, such that more experiences can be acquired collectively.

Regarding the above intentions of QA Lab-3, DG Lab is particularly interested in pragmatic aspect of rapid prototyping for summarization and evaluation. Thus, team DGLab is devoted to contributing a loosely coupled pipeline that keeps the effort as minimal as possible, by assembling highly accessible toolkits and data. Meanwhile, since the official result to date only examines long essay questions but not the short ones, this paper will also draw attentions toward it.

The rest of the paper is organized as follows: Section 2 describes the specification of long essay question. Section 3 and 4 present a system pipeline and an automatic evaluation metrics, respectively, for essay QA with minimal effort. Regarding team DGLab's work, Section 5 lists selected result and Section 6 discusses about it subsequently. Finally, Section 7 concludes.

2. LONG ESSAY QUESTION

The essay QA task of NTCIR QA Lab 3 classifies as long/complex questions and related shorter/simpler ones. The former requires multiple (typically 5 to 8) sentences (ranging from 225 to 270 words) including 8-10 designated keywords. The latter expects an answer essay in one or two sentences (normally 15 to 60 words) while some of them might be a factoid question. A gist of long essay questions and one of its short counterparts, from the question set P792W10 (The University of Tokyo, 2014), are listed as follows:

- Long essay (P792W10-1): … discuss the changes that Russian foreign policy had on the international situation throughout Eurasia from the Congress of Vienna to the end of the 19th century, noting how the western powers responded. … Limit your answer to 300 English words or less. Use each of the eight terms below at least once…
    - Afghanistan, Ili region, Primorye, Crimean War, Treaty of Turkmenchay, Berlin Conference (1878), Poland, Port Arthur
- Short essay (P792W10-2): The Byzantine Empire (Eastern Roman Empire) conquered many regions around the Mediterranean Sea during the 6th century reign of Emperor Justinian, but following his death, it gradually lost control of these lands. During this process, attacks by countries established by Turkish peoples had a tremendous impact on the history of the Byzantine Empire. Give an account of this in four lines or less.

Please kindly note that while the short essay question is mentioned, since this type has no official evaluation records from the organizer, the rest of the paper will be much more focused on the long essay question accordingly.

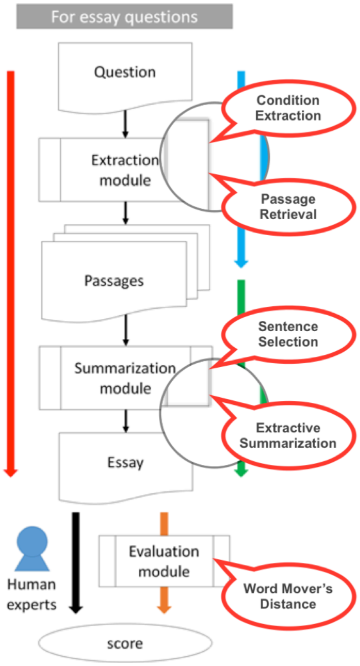

Figure 1. System Pipeline

3. PROPOSED SYSTEM

Figure 1, courtesy of the QA Lab 3 website[1], which illustrates the subtasks as modules for essay questions, is augmented to frame the pipeline of DGLab essay question-answering system at NTCIR-13 QA Lab-3. The pipeline comprises four stages, namely condition extraction, passage retrieval, sentence selection, extractive summarization. Regarding the subtasks, the former and the latter two stages correspond to the exaction and the summarization modules, respectively. While the abstract stages remain consistent, the decisions of corresponding implementations between Japanese and English essays are different and will be addressed accordingly when necessary.

While remaining aligned in the figure, the proposed system does not consult the evaluation module for question answering, since its purpose for the time being is to study correlations between automatic evaluation metrics and manual judgments by human experts. The evaluation method will be described separately in the next section.

3.1 Condition Extraction

Here the condition stands for both hard constraints such as answer length limit and soft ones like potential query terms. Condition Extraction stage is responsible for interpreting answer sheet XML files and compiling a list of categorized arguments for the downstream stages. The pseudo code below defines steps.

```
Map text by element/attribute name as type for each XML node
Extract answer length limit by regular expression
Categorize “keyword” type as required query term
Extract nouns from the rest descriptive type of texts
Categorize the above extracted nouns as optional query term
```

As for the descriptive texts, “grand_question,” “instruction,” “reference,” and “viewpoint” types of elements are all included. These types along with “keyword” and “answer length limit” will be referred as reserved terms in the rest of the paper. To extract nouns from them, the implementation only relies on coarse-grained part-of-speech. Words annotated with “名詞 (noun)” by MeCab[2] and “NN” by Stanford CoreNLP[3] are acquired for Japanese and English, respectively. Please note that in spite of NEologd[4] has been applied to MeCab to expand the vocabulary of named entity, for the tested questions at least, there is virtually no differences to the original MeCab dictionary, due to the fact that NEologd is mainly Internet and pop-culture oriented.

3.2 Passage Retrieval

To perform a comparable study to related work, and in hopes of utilizing the abilities of ordered window and pseudo relevance feedback, the chosen full-text search engine is Indri[5].

[1] http://research.nii.ac.jp/qalab/task.html, retrieved in Oct. 2017.

[2] http://taku910.github.io/mecab/, version 0.996

[3] https://stanfordnlp.github.io/CoreNLP/, version 3.7.0

[4] https://github.com/neologd/mecab-ipadic-neologd, 20170116

[5] http://www.lemurproject.org/indri.php, version 5.10

3.2.1 Source

For reproducibility and simplicity, passages are strictly from QA Lab organizer's data collection as follows:

- World history textbooks in Japanese
- All Wikipedia articles in Japanese
- Wikipedia articles in English.

In terms of the definition of passage, although paragraphs or ranged/overlapped sentences are often applied in factoid question-answering systems, this study follows the original document scope, in the spirit of being a minimal effort baseline. A heuristic of sentence selection to be described in the next section serves as a practical replacement of a predefined passage.

3.2.2 Indexing

Character-based index is adopted for Japanese without any special treatment such as lemmatization or stop-word filtering, while sentence segmentation by spaCy[6], Penn Treebank style tokenization by Stanford CoreNLP, and Krovetz stemming are sequentially applied for English. Indexed passages and their identifiers are stored in SQLite[7] databases for convenience.

3.2.3 Retrieval

Required and optional query terms are interpreted as Indri's ordered and unordered windows, respectively. For example, a text box below demonstrates the query for the long essay question P792W10-1, where "#combine()" and "#1()" resemble the concept of logic operators "OR" and "AND," respectively. To be more precise, "#1()" indicates that the terms inside must appear in the same word order within one-term distance. The layout is only for readability.

```
#combine(
   #1(Afghanistan) #1(Ili region) #1(Primorye) #1(Crimean War)
   #1(Treaty of Turkmenchay) #1(Berlin Conference 1878)
   #1(Poland) #1(Port Arthur)
)
```

Optional query terms are only applied to the short essay questions, for two arbitrary reasons: the absence of required query terms in the short essay questions, and a preparatory study in the long essay questions that has suggested their ineffectiveness and redundancy. In some cases of long essay questions, crucial information can be left out regrettably. This issue will be further examined latter in the section of discussion. The result set represents top-10 documents. The choice of 10-document threshold is empirical, based on observations of Indri score and speed.

3.3 Sentence Selection

Acting as an approximated validation, the sentence selection stage uses a heuristic defined as follows:

```
Segment sentences for each passage
Sort sentences by precision in terms of matched keyword
Compose passages by top-7 sentences unless 10 times longer
than answer length limit in total
```

Again, the 7-sentence threshold is empirical, while the 10-time total length limit is based on the QA Lab-3's submission format specification.

3.4 Extractive Summarization

With respect to answer length limit, Shuca[8][3] for Japanese and sumy's LexRank[9][1] for English are applied. The former one is modeled as redundancy and length conditioned Knapsack problem. The latter one prioritizes each sentence in a tf-idf and PageRank fashion. Despite the difference between those two approaches, both of them evaluate and then extract sub-texts based on their importance. Implementation-wise, Shuca requires JUMAN [10] and KNP [11] to analyze syntactic and semantic structures, while LexRank only utilizes surface patterns. Arbitrary passage truncation and Pointer-Generator Networks[4] have been tested as preliminary experiments. The former has been submitted as the secondary run of summarization subtask in Phase 2, while the latter has not because of lacking guaranteed convergence and sufficient training data in Japanese.

4. EVALUATION METHOD

The proposed method simply measures Word Mover's Distance[2] between gold standard nugget and system produced summary, yet another similarity-oriented metrics, based on Euclidean distance in terms of unit-normed word embeddings. The embeddings are generated from world history textbooks for Japanese and selected Wikipedia articles for English. The vector training and similarity estimating functions are both supplied by gensim[12]. The implementation of this study to-date doesn't involve any further refinement such as stop-word filtering.

5. RESULT

Team DGLab has participated Phase-2 and Research Run for all subtasks except the end-to-end one in Phase-2, due to the time constraint (mainly spent on the failed attempts of Pointer-Generator Networks). The organizer, however, is kind enough to officially list DGLab's summarization outcome for comparisons with other end-to-end systems. Hence it is reported here as well. To avoid unnecessary redundancy, detailed statistics that can be found in the organizer's overview paper are excluded. Specifically, for end-to-end and summarization subtasks, only human expert judgments for Phase-2 are listed. Subsequently, only averages of correlations to these human expert judgments are presented in this paper for evaluation method subtask.

5.1 End-to-end/Summarization Scores

Table 1 and 2 list scores graded by human experts for Japanese and English long essay questions, respectively. Their first columns indicate each prefix letter of question label as ID, except the last rows are sums of scores, to imitate a common practice of grading exams involving human subjects, which may or may not be the fairest choice of assessing performance of automatic systems. Another detail probably worth mentioning in advance is that since no participants receive positive scores for the question L, it will be excluded from the correlation coefficient estimation in the next subsection.

[6] https://spacy.io, version 1.7.2

[7] https://www.sqlite.org, version 3.18.0

[8] https://github.com/hitoshin/shuca, version 2016-01-15

[9] https://github.com/miso-belica/sumy, version 0.6.0

[10] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN, version 7.01

[11] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP, version 4.14

[12] https://radimrehurek.com/gensim/, version 2.1.0

Table 1. Scores for Japanese questions

	DGLab	Forst	IMTKU	KSU
B	3	5	3	3
C	0	1	0	1
G	0	0	0	1
L	0	0	0	0
P	12	3	0	4
Sum	15	9	3	9

Table 2. Scores for English questions

	CMUQA	DGLab	IMTKU	MTMT
B	-9	-8	-7	-4
C	-6	-5	-7	-15
G	-7	-7	-3	-17
L	-11	-7	-8	-19
P	-8	-8	-5	-12
Sum	-41	-35	-30	-67

5.2 Evaluation Method Coefficients

Table 3 and 4 present averages of Spearman's rho (ρ_{avg}) and Kendall's tau-b ($\tau\text{-}b_{avg}$) rank coefficients between the human judgments and the evaluation metrics by team per run, along with ROUGE-1 and nugget voting by participants as benchmarks, for Japanese and English, respectively.

Table 3. Coefficients for Japanese questions

	ρ_{avg}	$\tau\text{-}b_{avg}$
ROUGE-1	0.62	0.59
Nugget Voting	0.43	0.38
DGLab-2	0.34	0.30
Forst-1	-0.07	-0.05
Forst-2	0.40	0.36
tmkff-1	0.19	0.21

Table 4. Coefficients for English questions

	ρ_{avg}	$\tau\text{-}b_{avg}$
ROUGE-1	-0.26	-0.21
Nugget Voting	0.09	0.07
DGLab-2	-0.16	-0.07

To prevent from any confusion, coefficients of DGLab-1 (the only officially submitted run by team DGLab) are excluded, since an implementation error that renders the metrics meaningless has been revealed afterwards. It is also important to point out that DGLab-2 coefficients, despite being the corrected version of the proposed method, is merely a post-deadline reference, not a formal run submission.

According to the information provided by the organizer, coefficients of ROUGE-1 listed in Table 4 for English questions has applied stop-word filtering, while the Japanese counterpart in Table 3 is based on content word only. Those two treatments could be beneficial for the proposed Word Mover's Distance application, if not just for consistency and comparability, especially when stop-word filtering treatment is quite common in English usages of Word Mover's Distance.

6. DISCUSSION

While the performance varies for each question, team DGLab achieved 1st and 2nd places in Phase-2 for Japanese and English, respectively. Please note that for English questions, no teams have earned positive scores. Therefore the discussion is carried out with only Japanese questions. In particular, examples are drawn from the essay questions L792W10-1 (The University of Tokyo, 2010), since no participants gain any positive points on it in terms of expert score.

The question L792W10-1's instruction and keywords are presented below:

- … describe the role of the Netherlands and the Dutch people in world history, from the late Middle Ages to the modern day, when integration is extending beyond national lines…
    - Grotius, coffee, Pacific War, Nagasaki, New York, Habsburgs, Treaty of Maastricht, South African War

The submitted Japanese summary is roughly translated into English and partially listed by sentence with matched keywords indicated by underlines as follows:

1. Karl V united Netherlands 17 states under the House of Hapsburg's ruling.
2. The ports of New York and Boston prospered as slave trade port.
3. South African War is also known as Boer War.
4. … took effect the European Union Treaty (Treaty of Maastricht)… in 1993, and European Union (EU) established.
5. Roma Treaty… EC constitution in 1992 … Treaty of Maastricht (European Union Treaty) … in '93
6. With the start of Pacific War, not only Germany and Italy have declared war against the U.S… the Axis of Japan, Germany, and Italy… the U.S., the U.K., and Soviet Union…
7. … on December 8, 1941, Japan made a surprise attack on Hawaiian Pearl Harbor, and declared war against the U.S. and the U.K., so Pacific War (Asia-Pacific War) began.
8. However, when the Pacific War (Asia-Pacific War) began, … had to wait for end of the war to become independent.
9. In hamburger and coffee shops developed in a global scale…

It might be suffice to say that the summary has failed to be question-biased, since "Treaty of Maastricht" and "Pacific War" are referred multiple times without a coherent discourse structure. A more fundamental issue could be that only the first sentence mentions Netherlands, and the topic is hardly about its role in the world history.

The symptoms suggest several non-mutually exclusive causes that origin from the knowledge source, the extraction module, and the summarization module. It is likely that for the specific question, passages of such topic may be indirect/nonexistent in the source texts, or the applied query terms are insufficient to find relevant sentences, especially when Dutch-related information is only indicated outside the keywords. Based on preliminary studies, a conscious choice has been made to exclude optional query terms for long essay questions, and the above case shows that it is a design flaw that deserves more attentions. In fact, after revisiting the extracted passages, it has become clearer that almost (if not always) every mention of Dutch/Netherlands is at least one-sentence away from a sentence that contains the designated keywords. For example, the 3rd sentence of the submitted summary says

> *"South African War is also known as Boer War."*

While this sentence is actually right next to a sentence saying

> *"Boer War (Anglo Boer War) is the second war that the U.K. and the Dutch Boer (Afrikaner) fought for colonization in South Africa."*

Another related postmortem observation reveals that, for "South African War" as a surface pattern, it has been associated with passages from Wikipedia, not the Japanese textbooks. Sometime Wikipedia and textbooks both have relevant, but the former one often outranks the latter according to Indri or keyword-centered metrics, even though sentences from the latter one can be more concise and informative in terms of the given essay question. For example, the 2nd sentence of the submitted summary is from a long Wikipedia passage matches both "New York" and "coffee," but a short, low-rank passage from textbooks includes only "New York" yet describes its historical events involving "New Netherlands" and "New Amsterdam" in just one sentence:

> *"... the U.K. took the New Netherlands colony that the Netherlands opened, and seized its central New Amsterdam and renamed it New York."*

In retrospect, it could be much easier for the summarization module to pick up Dutch-related phrases.

With hindsight, on one hand, it could be beneficial if valid terms can be injected into a summarization system to produce question-biased summaries. On the other hand, from extraction to summarization, neither the heuristic sentence selection nor the sentence-importance based extractive summarization is capable of preserving discourse structures or even (necessary) synonyms, which leads to a bag of sentence rather than a qualified essay. Although it has been an inherent defect as expected, a semantics and pragmatics enabled summarization algorithm is nonetheless more desired.

7. CONCLUSION

The paper describes DGLab question answering system and automatic evaluation method at NTCIR-13 QA Lab-3 for Japanese university entrance exam on world history essay. Submissions of subtasks include extraction, summarization, and evaluation method, in Phase-2 and Research Run for both Japanese and English. Particularly on summarization subtask, the intended minimal effort system has received competitive performances for both Japanese and English, in terms of total scores graded by human experts. For the automatically answered essays to be able to actually pass the entrance exam, however, there still a long way to go.

8. ACKNOWLEDGMENTS

This work is supported by DG Lab Fund, which is managed by DG Daiwa Ventures, a partnership established by Digital Garage, Inc. and Daiwa Securities Group, Inc.[13]

9. REFERENCES

[1] Erkan, G. and Radev, D.R. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*. 22, (2004), 457–479.

[2] Kusner, M.J., Sun, Y., Kolkin, N.I. and Weinberger, K.Q. 2015. From Word Embeddings To Document Distances. *Proceedings of The 32nd International Conference on Machine Learning* (2015), 957–966.

[3] Nishikawa, H., Hirao, T., Makino, T. and Matsuo, Y. 2012. Text Summarization Model based on Redundancy-Constrained Knapsack Problem. *Proceedings of COLING 2012* (2012), 893–902.

[4] See, A., Liu, P.J., Brain, G. and Manning, C.D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, Canada, 2017), 1073–1083.

[5] Shibuki, H., Sakamoto, K., Ishioroshi, M., Fujita, A., Kano, Y., Mitamura, T., Mori, T. and Kando, N. 2017. Chronological and Geographical Measures for Evaluation of World History Essay QA in University Entrance Exams. *Proceedings of Open Knowledge Base and Question Answering Workshop at SIGIR 2017* (2017).

[6] 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. *Proceedings of NTCIR-12* (2016), 392–408.

[7] 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 NTCIR-13* (2017).

[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 NTCIR-11* (2014), 518–529.

[13] http://www.garage.co.jp/en/pr/2016/07/20160722.html, retrieved in Oct., 2017