IMTKU Question Answering System for World History Exams at NTCIR-12 QA Lab2

Min-Yuh Day[*1], Cheng-Chia Tsai[1], Wei-Chun Chuang[1], Jin-Kun Lin[1], Hsiu-Yuan Chang[1], Tzu-Jui Sun[1], Yuan-Jie Tsai[1], Yi-Heng Chiang[1], Cheng-Zhi Han[1], Wei-Ming Chen[1], Yun-Da Tsai[1], Yi-Jing Lin[1], Yue-Da Lin[1], Yu-Ming Guo[1], Ching-Yuan Chien[1], Cheng-Hung Lee[2]

[1] Department of Information Management, Tamkang University, New Taipei City, Taiwan,
[2] Sagacity Technology Co., Ltd.

myday@mail.tku.edu.tw, {petertsai0224, aristocratggc, saxphone30230, yuio798, donsaylazy52, st25662937, envychiang, peter09830961, will19940625, tydjason0221, pkjack9504, simon08074, sam5337882, lobster9160}@gmail.com, evan@stco.tw

ABSTRACT

In this paper, we describe the IMTKU (Information Management at Tamkang University) question answering system for Japanese university entrance exams at NTCIR-12 QA Lab-2. We proposed a question answering system that integrates natural language processing and machine learning techniques for Japanese university entrance exams at NTCIR-12 QA Lab-2. In phase-1, we submitted 6 End-to-End QA runs results for only English subtask for National Center Test for University Admissions and Secondary exams subtask. In phase-3, we submitted 7 End-to-End QA run results for English and Japanese subtask for Nation Center Exams and Secondary exams subtask. In NTCIR-12 QA Lab-2 phase-1, the IMTKU team total score achieved 31, 27 and 0 in the English subtask. In NTCIR-12 QA Lab-2 phase-3, the IMTKU team total score achieved 20, 20 and 14 in the English subtask, 24, 12 and 8 in the Japanese subtask and 31 in a combination run with KitAi.

Categories and Subject Descriptions

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

General Terms

Experimentation

Team Name

IMTKU

Subtasks

QA Lab-2 (National Center Exams English Version, Secondary Exams English Version, National Center Exams Japanese Version)

Keywords

IMTKU, NTCIR 12, QA Lab-2, World History, Question Answering, Machine Learning, University Entrance Examination, Essay Question, Answer Validation.

1. INTRODUCTION

IMTKU participated in NTCIR-12 QA Lab-2 National Center Test for University Admissions and Secondary exams in Japanese and English version from Japanese University entrance exams. NTCIR-12 QA Lab-2 totally has three phases, the English subtask will be done in two phases (Phase-1 and Phase-3). The Japanese subtask will be done in three phases (Phase-1, Phase-2 and Phase-3). We participated in Phase-1 and Phase-3. In Phase-1, we submitted 6 Question Analysis results (QA), 6 IR run results (RS), 6 End-to-End QA run results (FA) for only the English subtask for National Center Tests and Second-stage Examinations. In Phase-3 we submitted 6 Question Analysis results (QA), 12 IR run results (RS), 6 End-to-End QA run results (FA) and 1 Combination run results for English and Japanese subtask for National Center Exams and Secondary exams subtask. In this paper, we describe the tools and resources used in IMTKU QA Lab-2 question answering system.

Question Answering (QA) is a CLEF/TREC task of deciding a given question, whether the question answering returns a correct answer or not which is widely applied for many languages such as English, Russian, French, Japanese, Chinese, etc. QA-Lab is to provide a module platform for solving real-world entrance exam question by NTCIR-11 [5, 7].

QA-Lab is to solve real-world university entrance exam questions for world history. QA-Lab1 is the first pilot task in NTCIR-11. The world history questions are used from The National Center Test for University Admissions and Secondary exams including Japanese and English translations version. NTCIR-11 QALab1 question types are True/False questions, factoid questions, and a number of questions with short answer of Japanese characters [5, 7].

For instance, in a question answering sytem, an input question and answer shows as follows:

Question: 2012 年美國總統是誰？
(Who was the U.S. president in 2012?)
Answer: 巴拉克·歐巴馬
(Barack Obama)

The output of a question answering system is an answer for an input question.

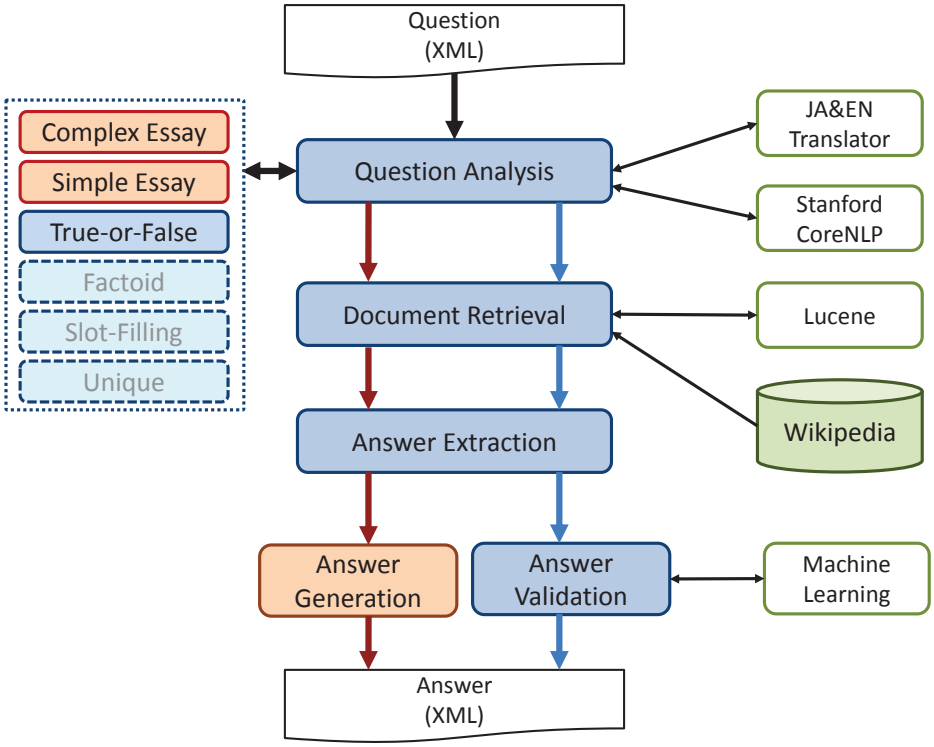

Figure 1. The System Architecture of IMTKU Question Answering System at NTCIR-12 QA Lab-2

QA-Lab provided the module structure of the original QA platform. The architecture is divided into 4 modules, question analysis, document retrieval, extraction of answer candidates, and answer generation [7]. The Question analysis module analyzes the types of questions and extracting question format. Document retrieval focused on search related documents using retrieval tools such as Apache Lucene and Lemur Indri. Answer Extraction means extracting answer candidates from the document retrieval using retrieved document or passage. Answer Generation focused on ranking the answer candidates based on the ranking score [7].
NTCIR-12 QA Lab-2 subtask is slightly different from NTCIR-11 QA-Lab1. In NTCIR-12 QA Lab-2, the organizers defined six question types such as Complex Essay (CE), Simple Essay (SE), Factoid (F), Slot-Filling (SF), True-or-False (TF) and Unique (U). Phase-1 is also slightly different from Phase-3. Phase-1 subtask does provide the Question Type Table, but Phase-3 doesn't provide Question Type Table. All participants need to select the question types or only one question type [6].

According to question types distinctions the evaluation can be different. The evaluation uses scores from National Center for University Admissions and other universities with Factoid, Slot Filling, True-or-False and Unique question type. Complex Essay and Simple Essay evaluation uses various of ROUGE and pyramid method.

The organization of this paper is as follows. Section 2 describes the system architecture. Section 3 describes the experimental results and analysis. Finally, we present discussions in Section 4 and conclude our work in Section 5.

2. SYSTEM ARCHITECTURE

Figure 1 shows the system architecture of IMTKU Question Answering System at NTCIR-12 QA Lab-2. The system architecture consists of five major components with dedicated modules for various exam question format. We describe the five major components, namely, question analysis, document retrieval, answer extraction, answer generation, and answer validation in this section.

2.1 Question Analysis

First, we used Stanford CoreNLP include Stanford NER and Stanford POS tagger added each word from raw dataset. Because these two tools helped analyze each word for tag and distinguished them by using splash (origin word/NER result/POS result). For example, through this tools analysis results for this word "Asia", we got the "Asia/LOCATION/NNP/".
Next, we classified question, according to question type and extracted each question and answer sentence to important keywords.

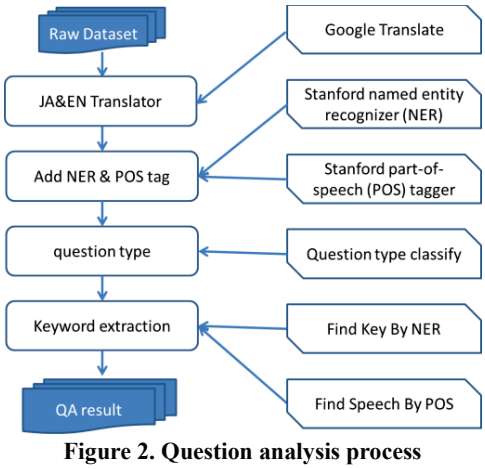

Figure 2. Question analysis process

Figure 2 shows that question analysis process. In question analysis, we classified question type about Simple Essay and Complex Essay to Document Retrieval, Answer Extraction and Answer Generation to deal with; Question type about True-of-False to Document Retrieval, Answer Extraction and Answer

Validation to deal with, but now we didn't well on Unique, Slot-Filling and Factoid.

```
<QUESTION_FORMAT_TYPE_DEF PARENT="ROOT" CHILDREN="E-C"
LABEL="Essay">E</QUESTION_FORMAT_TYPE_DEF>
  <QUESTION_FORMAT_TYPE_DEF PARENT="E" CHILDREN="E-C-K"
LABEL="Complex Essay">E-C</QUESTION_FORMAT_TYPE_DEF>
  <QUESTION_FORMAT_TYPE_DEF PARENT="E-C" CHILDREN="E-C-K-TR,E-C-
K-T" LABEL="Complex Essay with Keyword">E-C-
```

Figure 3. Sample Question format type set

Figure 3 shows that sample question format type set. NTCIR-12 QA Lab-2 official provided question format type set in Phase-1. But in phase-3 didn't provide question format type set.

Keyword extraction did extracted question content and extra content (underline with question), because of those content already had POS and NER tag. Therefore, we used those content to extract important word. We deleted interjection, verb to be from POS result but remained verb, adjective, Noun and NER result as keyword. If we handle Japanese version question, we used Google translate to translate Japanese to English.

2.1.1 Training

We used center exam (1997, 2001, 2003, 2005, 2007, 2009) and secondary stage examination (Tokyo) as training datasets of our system.

Table 1. Knowledge Type Table

Type of knowledge source	Knowledge_type
外部知識	KS
相片/圖片 理解	IC_P
地圖 理解	IC_M

```
<data id="D0" type="text">
    <label>A</label><br />Writing about trends among highly-educated people during the
Ming period, the Qing period scholar Zhào Yì states that from the <uText
id="U1"><label>(1)</label>Tang and Song periods onwards, most of those who excelled
in culture and the arts were those who had passed the Imperial examinations</uText>, but
in the <uText id="U2"><label>(2)</label>Ming period</uText>, there was a shift toward
figures outside the bureaucracy.The painter Tang Yin, who lived during the middle part of
the Ming period, can truly be described as a key figure from that transitional period. While
achieving outstanding marks in the Imperial examinations, he became embroiled in an
unfortunate incident; after the path to advancement was barred to him, he made his living
from painting in Suzhou, while living a carefree lifestyle.From the middle to the late Ming
period, a succession of artists and writers outside the bureaucracy emerged after culture
matured in cities, due to <uText id="U3"><label>(3)</label>the development of
commerce and industry, focused mainly on the Jiangnan region</uText>, with pictures and
publications coming to possess wide-ranging value as products.</data>
<question anscol="A1" answer_style="multipleChoice" answer_type="sentence" id="Q2"
knowledge_type="KS" minimal="yes">
<label>Question 1</label>
<instruction>
In relation to the underlined portion <ref comment="" target="U1">(1)</ref>, the figures
listed below are all people who passed the Imperial examinations in the Tang or Song
periods.From1-4 below, choose the one sentence that is correct in regard to the
person/people that it describes.
</instruction>
<ansColumn id="A1">1</ansColumn>
<choices anscol="A1" comment="">
<choice ansnum="1">
<cNum>(1)</cNum>Ouyang Xiu and Su Shi are writers representative of the Tang period.
</choice>
<choice ansnum="2">
    <cNum>(2)</cNum>Yan Zhenqing is a calligrapher representative of the Song period.
</choice>
 <choice ansnum="3">
    <cNum>(3)</cNum>Wang Anshi, who lived during the Song period, carried out
reforms called the New Policies (xin fa). </choice>
  <choice ansnum="4">
    <cNum>(4)</cNum>Qin Hui came into conflict with the party in favor of war,
concerning the relationship with the Yuan.  </choice>
  </choices>
</question>
```

Figure 4. Raw Datasets of NTCIR-12 QA Lab-2

2.1.2 XML dataset extraction

Datasets is required to reduce the noises in the raw data before analysis takes place. Therefore, in order to remove the noises in the raw datasets such as tags or labels. Figure 4 shows that NTCIR-12 QA Lab-2 raw Datasets. We extracted topic and content if topic and content are included in other related information, it extracted to analyze. For example, the system extracted dataset with underline (such as <uText>, <choice>, <instruction>, etc.) which question need.

In Question format type it classifies the questions to "question format type" by according to the topic is answer type and knowledge type. Answer type can distinguish six kinds of questions. Knowledge type can distinguish what information about external knowledge, picture, map and so on. The topic need as following:

Table 2. Answer Type Table

type of question	question_type
Unique-image	symbol-symbol
True/False-four description	sentence
Unique-time	o(symbol-symbol-symbol)
Slot-Filling	(symbol-term_other)*2
True/False-two description	(symbol-TF)*2
Factoid	others

Keywords play an important role in the process of IMTKU question answering system, because our system used keywords to search important sentences about history exams. Keywords of different topics by result and exclude some topic form with frequent term, interjection and be verb. extract Questions with POS Tagger and NER process it seizes Therefore, POS tagger and NER results is our keywords from training datasets. In three runs, we used different kinds of POS combinations and the results of NER to run those keywords.

2.1.3 POS Tagger

Figure 5 shows that an example of Stanford POS tagger and NER tools. Stanford POS Tagger is a part-of-speech tagging, it can read article or sentence and give part-of-speech of different words (e.g., verb, noun, adjective) and support other languages [9].

We used Stanford POS Tagger to analyze the part-of-speech of words in question. Stanford POS Tagger can divide the part-of-speech of words into thirty-six kinds of part-of-speech. After used Stanford POS Tagger to analyze topic and get a consequent and then we label the consequence behind each words.

Original:
Ouyang Xiu and Su Shi are writers representative of the Tang period.
Using POS Tagger and NER:
Wang/PERSON/NNP Anshi/PERSON/NNP ,/O/, who/O/WP lived/O/VBD during/O/IN the/O/DT Song/O/NN period/O/NN ,/O/, carried/O/VBD out/O/RP reforms/O/NNS called/O/VBD the/O/DT New/O/JJ Policies/O/NNS -LRB-/O/-LRB- xin/O/FW fa/O/FW -RRB-/O/-RRB- ./O/.

Figure 5. An Example of Stanford POS tagger and NER tools from NTCIR-12 QA Lab-2 center exam dataset.

2.1.4 Name Entity Recognition

Stanford CoreNLP also includes Stanford NER. The label of words which is marked by NER means name of object. It has the classifier used in NER and also many kinds of classifiers for user to choose.

We used Stanford NER to analyze the Named-entity of words in question. We used the classifier which can separate into seven kinds of category are location, organization, date, money, person, percent, and time.

2.1.5 JA&EN Translator

Figure 6 shows that an example of JA& EN translator. Because of our IMTKU system are all for English version, translate all content into English before analyze the Japanese version topic.

Japanese:
欧陽脩や蘇軾は，唐代を代表する文筆家である。
English (provided by organizer):
Ouyang Xiu and Su Shi are writers representative of the Tang period.
English (JA & EN Translator by Google Translate):
Ouyang Xiu and Su Shi is a writer representative of the Tang Dynasty.

Figure 6. An Example of JA & EN Translator

We used Google translate for Japanese center exam. we used the crawler to get the result and save back to our topic. After that, do the Question Analysis on the English topic which was translated from Japanese topic. The benefit is for free but it will need more time when translating. Figure 7 shows that NTCIR-12 QA Lab-2 of Question Analysis result.

```
<QUERY ID="1">
    <KEY_TERM_SET LANGUAGE="EN">
    <KEY_TERM RANK="1" SCORE="1">time</KEY_TERM>
    <KEY_TERM RANK="2" SCORE="0.95">Tsuda</KEY_TERM>
    <KEY_TERM RANK="3" SCORE="0.9">Umeko</KEY_TERM>
    <KEY_TERM RANK="4" SCORE="0.85">journey</KEY_TERM>
    <KEY_TERM RANK="5" SCORE="0.8">USA</KEY_TERM>
  </KEY_TERM_SET>
</QUERY>
```

Figure 7. NTCIR-12 QA Lab-2 of Question Analysis Result

2.2 Document Retrieval

Figure 8 shows that NTCIR-12 QA Lab-2 of Document Retrieval result Document Retrieval was important of IMTKU question answering system. We used Apache Lucene because it was important text search engine in Document Retrieval. The features of Lucene are cross-platform, typo-tolerant, ranked searching. All those data be processed to be more formal and then be stored Lucene indexes, for realizing convenient and quick search in QA systems [8]. Therefore, we used Apache Lucene in our Question Answering System.

Document retrieval got the keywords from question analysis result and Essay QA result. Essay QA result means Simple Essay and Complex Essay.

```
<DOCUMENT_SET>
  <DOCUMENT RANK="1" SCORE="0.5817159"
SOURCE_ID="https://en.wikipedia.org/wiki/Bruce_Lee"
SOURCE_ID_TYPE="WEB">Styles of Chinese martial arts;
  </DOCUMENT>
  <DOCUMENT RANK="2" SCORE="0.5720493"
SOURCE_ID="https://en.wikipedia.org/wiki/Dukkha"
SOURCE_ID_TYPE="WEB">Suffering between the periods of birth and
death:
  </DOCUMENT>
```

Figure 8. NTCIR-12 QA Lab-2 of Document Retrieval result

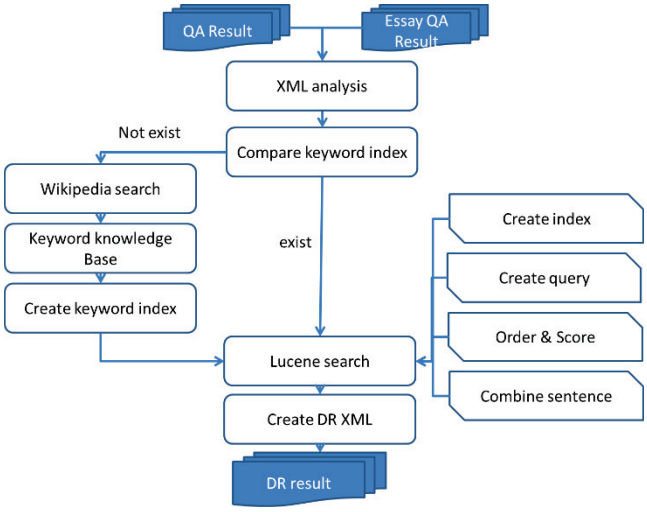

Figure 9. Document Retrievals process

Figure 9 shows that Document Retrievals process. First, if keyword didn't exist in our keyword knowledge base, we could search keyword sentence from Wikipedia. After that, we stored to Keyword knowledge base and create keyword index. If we compare keyword index had exist, we would use Apache lucene search. Next, the analyzer used to create index will be used on the sentence in the query string in knowledge base. Because the front and the behind sentence was related to keywords, we were combined 3 to 5 sentences. After that, we counted the each keyword sentences score and ordered those keyword sentence from high to low. Final, we were output IR run result.

2.3 Answer Extraction

Figure 10 shows that answer extraction process. Our input module of system separated in three parts, Question Analysis result, Document Retrievals result, official dataset. QA result include : (1) The keyword analyzed by questions. (2)Question type (3) Question Asking for (QA Not done). Our DR result includes keywords extracted from dataset, Wikipedia. And the usefulness of official dataset is select Answer Options.

We applied Main & Vice Key Processor in order to give the score to every sentences. MainKey is compose of the high frequency keyword in DR. ViceKey is the word in DR exclude MainKey. Then classify nodel separated ecah questions into six type: Factoid, Slot-Filling, choose the one sentence that is correct / incorrect, Choose the correct combination of "correct" and "incorrect" according to Question Type.

We applied all the tools in a handle model in system. It includes Score Answer, Score Sentence and Filling Answer in the Blank.

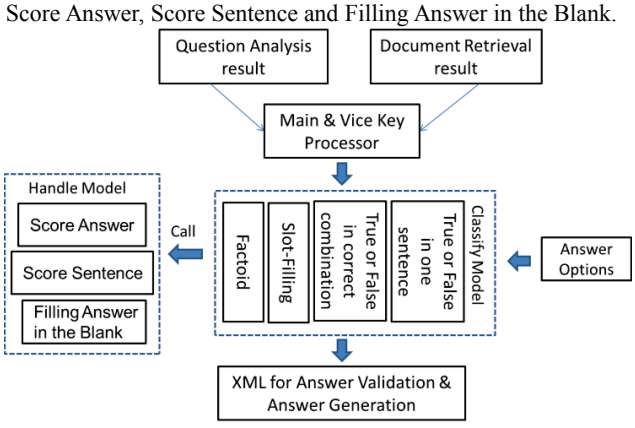

Figure 10. Answer Extraction Process

Score Answer is a special tool only for Factoid question. It may help system choose the right answer. Score sentence extract 20 sentences from DR, and then choose 2 sentences that have high relationship with the question to AG/AV. In order to improve the accuracy of answer, filling Answer in the Blank is a special tool for Slot-Filling. It send the question to AG/AV in order to help system choose a right answer. Then fill the answer in the blank. When every questions is done. The system was output a XML file that official need as the answer sheet. Figure 11 shows IMTKU question answering system of Answer Extraction result.

```
<keyList>
  <mainKey>work</mainKey>
  <mainKey>time</mainKey>
  <mainKey>letter</mainKey>
  <viceKey>1941</viceKey>
  <viceKey>son</viceKey>
</keyList>
<sentenceList>
  <sentence>The letter u ultimately comes from the Semitic letter Waw by way of the letter y</sentence>
  <sentence>In a 1941 letter to his son Michael, Tolkien recalled,</sentence>
</sentenceList>
<answerList>
```

Figure 11. Answer Extraction Result of IMTKU Question Answering System

2.4 Answer Validation

We used a machine learning approach by using LibSVM [1] for training our answer validation model. However, due to the technical problems for training the answer validation model for the formal runs, we used cosine similarity scoring function as the alternative answer validation in for the formal runs.

Cosine similarity implemented a discussion-bot to automatically answer students' queries by matching the reply posts from an annotated corpus of archived threaded discussions with students' query using cosine similarity [3].

Cosine Similarity Given a document retrieval sentence **S** and a candidate answer **A**, their cosine similarity weighted by inverse document frequency (idf) can be computed [2].

$$\mathrm{COS(S,A)}= \frac{\sum_{i=1}^{n}(S_i \times A_i)}{\sqrt{\sum_{i=1}^{n}(S_i)^2} \times \sqrt{\sum_{i=1}^{n}(A_i)^2}}$$

```
<sentenceList>
  <sentence>During this time, the Song court retreated south of the Yangtze River and established their capital at Lin'an  Although the Song Dynasty had lost control of the traditional birthplace of Chinese civilization along the Yellow River, the Song economy was not in ruins, as the Southern Song Empire contained 60 percent of China's population and a majority of the most productive agricultural land</sentence>
  <sentence>Southern Tang was conquered in 976 by the Northern Song Dynasty</sentence>
</sentenceList>
<answerList>
  <A>
    <answer>Ouyang Xiu and Su Shi are writers representative of the Tang period.</answer>
</A>
```

Figure 12. An Example of Answer Validation used dataset from Answer Extraction

Figure 12 shows that example of Answer Validation used dataset from Answer Extraction We got two sentences in sentence list from Answer Extraction. Next, we used sentenceList and answerList to calculate cosine similarity.

```
<combine>
  <sentence>In the 18th century, jesters had died out except in Russia, Spain and Germany</sentence>
  <answer>French troops, suffered a guerrilla war in Spain.</answer>
  <option>1</option>
  <dict>0.188982236504614</dict>
</combine>
```

Figure 13. An Example of Answer Validation Results

Figure 13 shows that an example of answer validation results. Final, we selected highest score for answer from option 1 to 4.

2.5 Answer Generation

We divided this step into two parts. The first part is about data reading that in order to obtain questions data and the result of document retrieval. We used Stanford coreNLP for analysis every word POS. Otherwise, we classify the result into six parts. The six attributes are Date, Location, Organization, Money, People, Percentage, Time, and we mark these words as keywords. The second part is about answer generation.

```
-<question id="q01" knowledge_type="RT" answer_type="sentence" answer_style="description_limited" minimal="yes">
<label>[1]</label>
<data id="d01" type="text">Throughout history, while contacts and exchanges between different cultures sometimes gave rise to conflict, such interaction also contributed significantly to diversification and transformation of cultures and lifestyles. For example, as the Arab-Islamic cultural sphere expanded, starting in the 7th century, different cultures were introduced from newly acquired territories and adjacent regions, resulting in further development. Those cultures, in turn, exerted influence on other regions.</data>
-<instruction>
Write a brief essay, within 17 lines, on those movements involving the Arab-Islamic cultural sphere up to the 13th century in answer section (A). Be sure to use all eight
<ref target="d02">keywords</ref>
shown below at least once, and underline those
<ref target="d02">keywords</ref>
</instruction>
<data id="d02" type="text">India, Abbasid Caliphate, Avicenna, Aristotle, medicine, algebra, Toledo, Sicily island</data>
</question>
```

Figure 14. An Example of Essay question

Figure 14 shows example of essay question. Our system output an XML file that according to frame that provided by the official dataset. The system chooses a proper keyword that fit the demand of question in the KEY_TERM_SET from the file of document retrieval. Stanford coreNLP is a toolkit that developed by Stanford university. It can analysis the part of speech of every word. However, due to the technique problem, we cannot reply the right answer.

3. EXPERIMENTAL RESULTS AND ANALYSIS

We participated Phase-1 and Phase-3, we conduct several experiments using various datasets (national Center Test for University Admissions and Secondary exams) to train and test models, as well as different results.

3.1 Phase-1

In this section, we describe the Phase-1 and each question format correct rate. We also present the results of formal runs.

In National Center Test, we submitted 3 runs "Center-1999-Main-WorldHistoryB_IMTKU_EN_QA_01, 02 and 03", question_format_type_set_def and answer_type_set_def for Question Analysis results. We submitted 3 runs "Center-1999-MainWorldHistoryB_IMTKU_EN_RS_01, 02 and 03" for IR run results. We submitted 3 runs "Center-1999-Main-WorldHistoryB_IMTKU_EN_FA_01, 02 and 03" for End-to-End QA run results.

In Second Stage Examination, we submitted 3 runs "D792W10_IMTKU_EN_QA_01, 02 and 03" for Question Analysis results. We submitted 3 runs "D792W10_IMTKU_EN_RS_01, 02 and 03" for IR run results. We submitted 3 runs "D792W10_IMTKU_EN_FA_01, 02 and 03" for End-to-End QA run results.

In Phase-1, the National Center Exams English subtask, system was evaluated by using scores from National Center for University Admissions, the Second-Stage Examination English subtask, two kinds of "Complex Essay" and "Simple Essay" are used for evaluation: ROUGE and pyramid method using nuggets.

We list the summary of IMTKU FA runs result for QA Lab-2 National Center Test and Second-stage Examination English subtask in Table 3, 4, 5, 6. Table 3 shows that result of multiple choice questions in Phase-1, the best average of our submitted score is 0.756. Table 4 shows that result of free description questions in Phase-1, the ROUGE-1 score is 0.0326, ROUGE-2 score is 0.00505. Table 5 shows that the Correct, Incorrect, Unanswered and score of IMTKU National Center Test 3 runs, the best score of our submitted FA runs for National Center Test English subtask is 31, which is "Center-1999--Main-World-HistoryB_IMTKU_EN_FA_01". The worst score of our submitted FA runs for National Center Test English subtask is 0. Lead to this result because our system appears systematic error. Table 6 shows that Correct, Incorrect, Unanswered and score of IMTKU Second-stage Examination 3 runs.

Table 3. Result of multiple choice questions in Phase-1

RUN	Correct rate	Total score	Average score
1	0.293	**31**	**0.756**
2	0.244	27	0.659
3	0.000	0	0.000

Table 4. Result of free description questions in Phase-1

RUN	Correct rate	ROUGE-1 score	ROUGE-2 score
1	0	**0.0326**	**0.00505**
2	0	0.00833	0
3	0	**0.0326**	**0.00505**

Table 5. Correct, Incorrect, Unanswered and score of IMTKU National Center Test 3 runs.

RUN	CORRECT	INCORRECT	UNANSWERED	Total	SCORE
1	12	19	10	12/41	**31**
2	10	21	10	10/41	27
3	0	0	41	0/41	0

Table 6. Correct, Incorrect, Unanswered and score of IMTKU Second-stage Examination 3 runs.

RUN	CORRECT	INCORRECT	UNANSWERED	WITHHOLDING	SCORE
1	0	8	22	1	0
2	0	8	22	1	0
3	0	8	22	1	0

3.2 Phase-2

We did not participate Phase-2 because Phase-2 is a Japanese only Subtask.

3.3 Phase-3

In this section, we describe the Phase-3 and each question format correct rate.

In National Center Test, we submitted 3 runs "Center-2011-Main-WorldHistoryB_IMTKU_EN_QA_01, 02 and 03" and "Center-2011-Main-WorldHistoryB_IMTKU_JP_QA_01, 02 and 03" for Question Analysis results. We submitted 3 runs "Center-1999-MainWorldHistoryB_IMTKU_EN_RS_01, 02 and 03" and "Center-1999-MainWorldHistoryB_IMTKU_JP_RS_01, 02 and 03" for IR run results. We submitted 3 runs "Center-1999-Main-WorldHistoryB_IMTKU_EN_FA_01, 02 and 03" and "Center-1999-Main-WorldHistoryB_IMTKU_JP_FA_01, 02 and 03" for End-to-End QA run results.

In Combination Run, we submitted run "Center-2011-Main-WorldHistoryB_KitAi_IMTKU_EN_FA_01".

In Second Stage Examination, we only submitted 6 runs "M792W10_IMTKU_EN_RS_01, 02 and 03" and "M792W10_IMTKU_JP_RS_01, 02 and 03" for IR run results.

In Phase-3, the evaluation is same as Phase-1, using scores from National Center for University Admissions. Question type table did not provided in Phase-3.

We list the summary of IMTKU runs result for National Center Test and Second-stage Examination in Table 7, 8, 9, 10. Table 7 shows that result of multiple choice questions in Phase-3, the best average score is 0.667.

Table 8 shows that the best score of our submitted FA runs for National Center Test English subtask is 20, which is "Center-2011-Main-WorldHistoryB_IMTKU_EN_FA_01" and "Center-2011-Main-WorldHistoryB_IMTKU_EN_FA_02". Table 9 shows that the best score of our submitted FA runs for National Center Test Japanese subtask is 24, which is "Center-2011-Main-WorldHistoryB_IMTKU_JP_FA_03". Table 10 shows that result of multiple choice questions for combination run in Phase-3. Table 11 shows that the best score of our submitted Combination run with KitAi for National Center Test English subtask is 34.

Table 7. Result of multiple choice questions in Phase-3

Run	Lang.	Correct rate	Total score	Average score
RUN01	EN	0.194	20	0.556
RUN02	EN	0.194	20	0.556
RUN03	EN	0.139	14	0.389
RUN01	JA	0.222	24	**0.667**
RUN02	JA	0.0833	8	0.222
RUN03	JA	0.222	24	**0.667**

Table 8. Correct, Incorrect and score of IMTKU National Center Test English 3 runs

EN	CORRECT	INCORRECT	SCORE
RUN01	7	29	**20**
RUN02	7	29	**20**
RUN03	5	31	14

Table 9. Correct, Incorrect and score of IMTKU National Center Test Japanese 3 runs

JA	CORRECT	INCORRECT	SCORE
RUN01	4	32	12
RUN02	3	33	8
RUN03	8	28	**24**

Table 10. Result of multiple choice questions for combination run in Phase-3

TeamID	Comb	Correct rate	Total score	Average score
IMTKU	KitAi	0.333	34	0.944

Table 11. Correct, Incorrect and score of IMTKU with KitAi National Center Test English Combination run

Combination	CORRECT	INCORRECT	Total	SCORE
RUN with KitAi	12	24	12/36	**34**

4. DISCUSSIONS

In order to evaluation the performance of IMTKU question answering system, we participated two phases (Phase-1 and Phase-3) of QA Lab-2 subtask. We used the train dataset of center exam (1997, 2001, 2003, 2005, 2007, 2009) and secondary stage examination (Tokyo) from organizers to train and test our question answering system.

In Phase-1, the results show that the best performance score of IMTKU question answering system in English subtask is 31 (FA run01). However, a systematic error of XML output issue was found in our submitted formal run (FA run03) and result in a 0 score.

However, in Phase-3, the results show that the best performance score of IMTKU question answering system in Japanese subtask is 24 "FA run03" is 24. There score is difference between the two. Although the score of Phase-1 is lower than the score of Phase-3, the score of Phase-3 with Combination is higher than the score of our system. Combination run means making a system answer using other system's result.

Figure 15 shows an example of answer format. IMTKU QA system aims to handle all question types; however, we could not resolve some question types well, such as Unique, Factoid and Slot-Filling. Specifically, we could not resolve particular answer format, such as symbol-symbol, o(symbol-symbol-symbol), (symbol-term_other)*2, and (symbol-TF)*2. In addition, although there are two types of knowledge format, namely, external knowledge and map comprehension. We used Wikipedia only from external knowledge and didn't utilize map comprehension.

```
o(symbol-symbol-symbol):
(1) a→b→c (2) a→c→b(3)b→a→c(4)b→c→a
symbol-symbol:
(1)(A)－a (2)(A)－b (3)(B)－a (4)(B)－b
(symbol-term_other)*2:
(1)(A)- Huguang(B)- Gongsuo (guild)
(2)(A)- Huguang(B)- Zujie (concession)
(3)(A)- Suhu (Jiangzhe)(B)- Gongsuo (guild)
(4)(A)- Suhu (Jiangzhe)(B)- Zujie (concession)
(symbol-TF)*2
(1)a- Correctb- Correct (2)a- Correctb- Incorrect
(3)a- Incorrectb- Correct (4)a- Incorrectb- Incorrect
```

Figure 15. An Example of Answer Format

5. CONCLUSIONS

In this paper, we proposed a question answering system using a hybrid approach that integrate natural language processing and machine learning techniques for Japanese university entrance exams at NTCIR-12 QA Lab-2. In phase-1, we submitted 6 End-to End QA run results only English subtask for National Center Test for University Admissions and Secondary exams subtask. In phase-3, we submitted 7 End-to End QA run results by English and Japanese subtask for Nation Center Exams and Secondary exams subtask. In NTCIR-12 QA Lab-2 phase-1, IMTKU team total score achieved 31, 27 and 0 in the English subtask. In NTCIR-12 QA Lab-2 phase-3, IMTKU team total score achieved 20, 20 and 14 in the English subtask, 24, 12 and 8 in the Japanese subtask and 31 scores in combination run with KitAi.

The contribution of this study is that we proposed the IMTKU Question Answering System focus on NTCIR-12 QA Lab-2 National Center Test and Second-Stage Exam. We integrated natural language processing tools and resources in the IMTKU Question Answering System with the capability for resolving the National Center Test for University Admissions and Secondary exams in Japanese and English

6. ACKNOWLEDGMENTS

This research was supported in part of TKU research grant and Ministry of Science and Technology. We would like to thank the support of IASL, IIS, Academia Sinica, Taiwan.

7. REFERENCES

[1] C. Chang and C. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), pp. 27. 2011.

[2] G. Cong, L. Wang, C. Lin, Y. Song and Y. Sun. Finding question-answer pairs from online forums. Presented at Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008.

[3] D. Feng, E. Shaw, J. Kim and E. Hovy. An intelligent discussion-bot for answering student queries in threaded discussions. Presented at Proceedings of the 11th International Conference on Intelligent User Interfaces. 2006.

[4] Y. Kano. Solving history problems of the national center test for university admissions. Presented at Proceedings of the Annual Conference of JSAI. 2014.

[5] K. Sakamoto, H. Matsui, E. Matsunaga, T. Jin, H. Shibuki, T. Mori, M. Ishioroshi and N. Kando. Forst: Question answering system using basic element at NTCIR-11 QA-lab task. Presented at NTCIR. 2014.

[6] H. Shibuki, K. Sakamoto,Y. Kano, T. Mitamura, M. Ishioroshi, T. Mori, N. Kando NTCIR-12 QA-Lab Task second Pilot.

[7] H. Shibuki, K. Sakamoto, Y. Kano, T. Mitamura, M. Ishioroshi, K. Y. Itakura, D. Wang, T. Mori and N. Kando. Overview of the NTCIR-11 QA-lab task. Presented at Ntcir. 2014.

[8] S. Song, Y. Meng, Z. Zheng and J. Sun. A feature-based classification technique for answering multi-choice world history questions. 2015.

[9] Stanford University, California. The Stanford Natural Language Processing Group. Stanford CoreNLP – a suite of core NLP tools. July 20, 2015 http://stanfordnlp.github.io/CoreNLP/

[10] M. Steinbach, G. Karypis and V. Kumar. A comparison of document clustering techniques. Presented at KDD Workshop on Text Mining. 2000.

[11] 2015 Academic Year Secondary Examination: World History (NTCIR12 QALab-2 task's Sample Questions), http://research.nii.ac.jp/qalab/sample_question/center_test_world_history_B_sample_question_en.html