Question Answering Challenge (QAC-1)
An Evaluation of Question Answering Task at NTCIR Workshop 3

Jun'ichi FUKUMOTO
Ritsumeikan University
fukumoto@cs.ritsumei.ac.jp

Tsuneaki KATO †
University of Tokyo†
kato@boz.c.u-tokyo.ac.jp†

Fumito MASUI ‡
Mie University ‡
masui@shiino.info.mie-u.ac.jp ‡

Abstract

In this paper we describe a question answering task, called Question Answering Challenge (QAC), and its first evaluation (QAC1). This was carried out as a task of NTCIR Workshop 3 in October 2002. In QAC, we aimed to encourage the development of practical QA systems in a general domain and focus on research of user interaction and information extraction. Developments of evaluation method of question answering system and information resources for evaluation were also purpose of QAC.

We have defined three kinds of task in QAC, which require five possible answers (Task 1), only one answer (Task 2) and one answer to related question (Task 3). We have prepared 200 questions for Task 1 and Task 2 and 40 questions for Task 3 at Formal Run and about 900 questions for additional run. We have conducted Dry Run and Formal Run evaluation and finally there are sixteen participants (two of them are from task organizer) at QAC1.

Keywords: *Question Answering, Information Extraction, Information Retrieval, user interaction, evaluation tools*

1 Introduction

Question Answering Challenge (QAC) was carried out as the first evaluation task on question answering of NTCIR Workshop 3 [2]. Question answering in open domain is a task to obtain appropriate answers to given domain independent questions written in natural language from a large corpus. The purpose of QAC is to encourage the development of practical QA systems in open domain and focus on research of user interaction and information extraction. Developments of evaluation method of question answering system and information resources for evaluation are also the purpose of QAC.

QAC was proposed in the last NTCIR Workshop 2 [1]. We had 20 organizing committee and hold four meetings to discuss evaluation method and some other problems on QA. We have also opened a web site for QAC at http://www.nlp.cs.ritsumei.ac.jp/qac/ in Japanese and English and started mailing lists.

2 Question Answering task

In order to evaluate QA technologies, there are several technical aspects to extract answer expressions from knowledge sources. Question type is one aspect of QA system evaluation. 5W1H type questions, which are question sentences using interrogative pronouns such as who, where, when, what, why and how, are typical one. In QA task, the points of question sentences are defined as a noun or noun phrase which indicates person names, organization names, names of various artifacts, money, size, date and so on. Moreover, their related information can be considered as answer candidates: as for person names, their affiliations, age, status name will be answer, and as for organization names, their annual profit, established year and so on.

Another aspect is related to how many answer expressions exist in knowledge sources. In the TREC QA task [3] [7] [4], they assumed there is only one answer for a question. However, there are multiple answers or no answer questions in general. This aspect makes development of QA system difficult. If multiple answers are assumed, system has to check all the answer candidates very carefully. If there is only one answer, system will choose highest priority answer from answer candidates with some priorities.

User interaction technology leads to actual interaction between computer and person. In actual QA between people, there will typically be several interactions in order to confirm the intention of the questions and so on. Information extraction that works in general domain is also an important technology in order to realize real QA system.

Answer text retrieval is an essential technology for QA system. In the first stage of question answering, several target texts will be retrieved for answer extraction using several key words in a given question sentence. The longer the sentence is, the more information exists for text retrieval. However, there are some

cases that several meaningless key words are embedded in question sentence.

3 Task Definition of QAC1

We will briefly describe the task definition of QAC1. As for target documents, we used two years of Japanese newspaper articles (1998 and 1999), Mainichi Newspaper articles. In QAC1, a question used for evaluation is a short answer question and an answer will be an exact answer which consists of a noun or noun phrase which indicates person names, organization names, name of various artifact, money, size, date and so on. These types are basically from Named Entity (NE) element of MUC [6] and IREX [5] but are not limited to NE elements.

In order to get answer, system can use other information sources such as an encyclopedia, thesaurus, corpus and so on. That is, it is permitted that answer expressions does not exist in newspaper articles. Moreover, paraphrased expressions are also permitted as answer expressions. However, why such answer expressions are correct should be justified by contents of newspaper articles.

We do not assume existence of one answer to a given question. That is, there is a case that there is no answer object in documents to a given question. Also, if there is multiple answer objects in a given documents, system has to respond all the possible answers as a list form.

We will give one or more follow-up questions to the first question. In Japanese, there will be ellipses in the follow-up question. For example, if the first question is a question of person name and the second question is a question of his/her affiliations, and so on.

Paraphrasing is also one aspect of QA system evaluation. In a target text, an answer expression may exist in other expression. In this case, system has to recognize paraphrased expression as the same one of original one. Otherwise, some expressions of a question sentence exist in paraphrased ones in a target text. In order to retrieve such a text to extract answer expression, identification of the same concept in various expressions is important technology.

3.1 Definition

According to the above outline of task definition, we have introduced three tasks in QAC1. The current version of QAC task definition is presented as follows:

- Task 1

  System extracts five possible exact answers from documents in some order. The inverse number of the order, Reciprocal Rank (RR), is the score of the question. For example, if the second answer is correct, the score will be half. The highest score will be the score of the question. If there are several correct answers, system will return one of them.

  For example, questions of this task are presented in the following. A question consists of QID (QAC1-1001-01) and question sentence (English translation is shown in parentheses.).

  > QAC1-1001-01: “2000 年 101 日に合併すみことが決まった通信三社はどこですか。(Which three telecommunications companies decided to merge on October 1, 2000?)”

  For the question, correct answers are “DDI”, “IDO” and “KDD”. System has to respond either of them.

  > QAC1-1002-01: “広辞苑第五版はいつ発売さみましたか。(When was the fifth edition of the Kojien Japanese dictionary published?)”

  For the question QAC1-1002-01, correct answer is “1111 (November 11)”. If there is an expression “昨 (yesterday)” in the article dated in Nov. 12, this answer will be correct. In QAC, relative expression of date is permitted, however, system has to give evidence that the answer was extracted from the particle in this case.

- Task 2

  Task 2 uses the same question set of Task 1 but evaluation method is different. System extracts only one set of answers from documents. If all the answers are correct, full score will be given. If there are several answers, system has to return all the answer. If there are some wrong answers, this will be penalty of the score. Average F-Measure (AFM) is used for evaluation of Task 2.

- Task 3

  This task is an evaluation of a series of questions or follow-up question in other words. A related question is given to a question of Task 2. There will be ellipsis or pronominalized elements in follow-up questions.

  For example, questions of this task are presented in the following. Question “QAC1-3011-02” is the follow-up question of question “QAC1-3011-01”. The “-02” means the first follow-up question of the main question, although there is only one follow-up question in the current task definition.

  > QAC1-3011-01: “久石譲が音楽を担当した宮在駿監督の映画は何ですか。(Joe Hisaishi was a music director for which of Hayao Miyazaki ’s films?)”

QAC1-3011-02: “北餅武の映画は何ですか。(What is the name of the film directed by Takeshi Kitano?)”

3.2 Support information

System is required to return support information for each answer of the questions, although it is optional. In the current definition, we assume the support information as one document ID which will be evidence of the replied answer.

4 Question development for evaluation

For QA evaluation, it is necessary to prepare a variety of questions which require a product name, title of novel or movie, numeric expression and so on. We have developed about 1200 questions in various question types that sometime include paraphrasing. Moreover, all the task participants are required to submit about twenty questions until Formal Run. Some of them will be used for the evaluation and others will be open as test corrections of QA data. The detail number of questions is summarized in Table 1.

Table 1. Prepared questions for QAC1

developer	number
task organizer	1202
task participants	200
Total	1402

5 Evaluation Method

5.1 Task 1

System extracts five answers from documents in some order. The inverse number of the order, Reciprocal Rank (RR), will be the score of the question. For example, if the second answer is correct, the score will be 1/2. The highest score of the five answers will be the score of the question. If there are several correct answers of a question, system might return one of them, not all of them. Mean Reciprocal Rank (MRR) is used for evaluation of task1. If n set of answers are correct, Mean Reciprocal Rank (MRR) can be calculated as follows:

$$MRR = \frac{\sum_{i=1}^{n} RR_i}{Q} \quad (1)$$

$$RR_i = \frac{1}{Rank} \quad (2)$$

For example, the following question “QAC1-1001-01” has an answer of three companies such as DDI, IDO and KDD.

QAC1-1001-01: “2000 年 101 日に合併すみことが決まった通信三社はどこですか。(Which three telecommunications companies decided to merge on October 1st, 2000?)”
Correct answer: DDI, IDO, KDD

Three kinds of answer evaluation are presented as follows:

- Response1: NTT, IDO, AT&T, NII, KDD
  RR=0.5
- Response2: AT&T, BT, DDI, IDO, KDD
  RR=0.33
- Response3: DDI, AT&T, BT,NII, Docomo
  RR=1.0

The underlined answers are correct ones. In Response1, system returned five answers in the above order and the second one and fifth one are correct. Therefore, RR value of the best answer (second one) will be score of this answer. In Response2, the third, fourth and fifth answers are correct, and then RR value will be 0.33. In Response3, only the first answer is correct, then RR value will be 1.0.

5.2 Task 2

System extracts only one set of answers from documents. If the system’s answer is correct, the score will be given. If there are several answers, a system has to return all the answer. Mean F-Measure (MF) is used for evaluation of Task 2. The scores are calculated in the following formula with assuming A as the number of correct answers, A_{sys} as the number of answers that the user’s system output, and A_{cor} as the number of correct answers that the user’s system output. Q and $Rank$ is assumed as the number of questions and rank of answer, respectively.

$$Recall = \frac{A_{cor}}{A} \quad (3)$$

$$Precision = \frac{A_{cor}}{A_{sys}} \quad (4)$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (5)$$

For example, system responses for the same question of Task1 “QAC1-1001-01” and their evaluations are presented as follows:

- Response1: NTT, IDO, AT&T, KDD
  P=2/4, R=2/3, F=0.57
- Response2: IDO, 日本移動通信 (IDO), KDD
  P=2/3, R=2/3, F=0.67

The underlined answers are correct ones as well as the above example. In Response1, system returned four answers for the question and the second and third ones are correct. Therefore, two of four answers are correct, and then precision will be 2/4. Also, two of three correct answers are detected, and then recall value will be 2/3. In Response2, correct answers are the first and third one, and then precision will be 2/3. Also, two of three correct answers are detected, and then recall value will be 2/3. In this case, the second one is the same organization of the first one, but the same elements will be ignored even if it is the same element.

In Task 1 and Task2, there is a case of no answer question. If there is no answer question and a system give no answer, the score of this question will be 1.0 (F-measure). But, if a system gives some answer for such no answer question, the score will be zero.

5.3 Task 3

This task is the evaluation of a series of questions. System has to return all the possible answers for a main question and its follow-up question. Score will be given only for the follow-up question in the same scoring method of Task 2, that is MF.

5.4 Scoring Tool

We have developed a scoring tool, written in Perl language, to help participants' evaluation. This tool can check whether answers of a system are correct or not by comparing the correct answer and the output. The tool can show each answer evaluation and some statistics of a task.

The scoring tool can provide all the result of answer checking to get information about whether each answer is correct or not. Figure 1 shows a part of a sample output.

```
QAC1-1020-01:  インド (India) , インドネシア (Indonesia) , タイ (Thailand) ×, 米 (USA) ×, フランス (France) ×
QAC1-1021-01:  φ ○
```

Figure 1. Sample output of scorer (answer check)

If the answer is correct, " (maru)" goes with the answer, otherwise, "× (batsu)". The "φ (phi)" (a symbol for phi) in the file means that the system output no answer in the question. The " (maru)" is given if and only if there is no answer in the correct answer set.

As for statistical results, this tool calculates the sum of correct answers and MRR for Task 1. For Task 2 and 3, this tool calculates the sum of correct answers and the mean F-measure. Figure 2 shows an sample output of this tool.

Task1 Results: 35.0 marks out of 200.0 in TASK1
Average score: 0.175

Question	Answer	Output	Correct
200	272	729	38
Recall	Precision	F-measure	MRR/MF
0.139	0.521	0.759	0.175

Figure 2. Sample output of scorer (score)

The first line summarizes the results and statistics for an input results and the following lines show the details of the score. "Question" is the total number of questions in the task and "Answer" is the number of different answers of questions. "Output" is the number of answers of the input data and "Correct" means the number of correct answers of the input data.

The detail usage of Scoring Tool is presented in Appendix B.

6 Task Participants

In QAC1, there were fourteen active participants. Task participation of each participant is shown in Table 2. The symbol "⋆" indicates that the team submitted one result to the task. Two symbols means two kinds of submission of results. The last two participations in italic are from QAC task organizers and are not included official score of the results of QAC1.

7 Runs for Evaluation

7.1 Description of Formal Run

We have conducted QAC Formal Run in the following schedule and tasks.

- Date of task revealed: Apr. 22, 2002 (Mon.) AM (JST)
- The result submission due: Apr. 26, 2002 (Fri.) 17:00 (JST)
- Number of questions:

  Task 1: 200
  Task 2: 200 (same as Task 1)
  Task 3: 40 (follow up questions of Task 1 questions)

[1]Participant name is taken from affiliation of the first author of presented paper.

Table 2. Task participation

participants name[1]	Task		
	1	2	3
Communication Research Laboratory	⋆	⋆	⋆
Kochi Univ. of Technology		⋆	
Matsushita Electric Ind.	⋆	⋆	
NTT Corp.	⋆⋆	⋆	⋆
NTT DATA Corp.	⋆⋆	⋆	⋆
Nara Institute Science and Technology	⋆	⋆	
National Institute of Advanced Industrial Science and Technology	⋆	⋆	⋆
New York Univ.	⋆		
Oki Electric Ind.	⋆	⋆	⋆
POSTECH	⋆		
The Graduate Univ. for Advanced Studies	⋆	⋆	
Toyohashi Univ. of Technology	⋆	⋆	⋆
Univ. of Tokyo	⋆	⋆	
Yokohama National Univ.	⋆		
Mie Univ.	⋆	⋆	
Ritsumeikan Univ.	⋆	⋆	

7.2 Additional QA runs

Task participants are required to evaluate about 900 questions in order to make more evaluation and develop better QA test collections. This was conducted after the Formal Run in the following schedule.

- Delivery of questions: May. 13, 2002 (Mon.)
- Submission due: May. 24, 2002 (Fri.)
- Submission format: same as Formal Run Formats

The submitted results were pooled and will be delivered after evaluation.

8 Results and Discussion

8.1 Task Analysis

In this subsection, we give a summary of the results of QAC formal run by each task. Throughout this chapter, system IDs the participants named their own systems are used in the figures; some of them are abbreviated because of the space problem.

Task 1

Fifteen systems participated in the Task 1. The accuracies of the participated systems achieved in the mean reciprocal rank (MRR) are depicted in Figure 3. The most accurate system achieved 0.61 in the MRR. This system returned the correct answer in the first rank to more than half of the questions, and in up to the fifth rank to more than three forth.

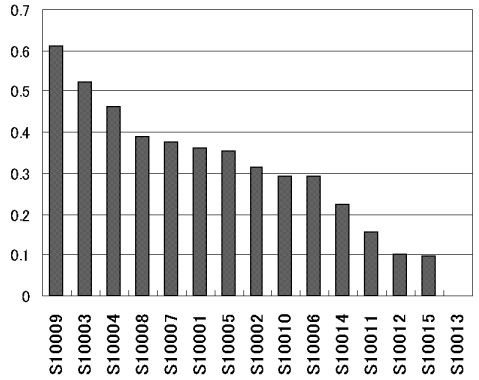

Figure 3. MRR of participant systems in Task 1

In addition to by the MRR measure, we tried evaluating the systems by two other measures. The first is the ratio that a system answered correctly in the first rank. The second is the ratio that a system answered correctly in up to the fifth rank. Those two measures show very small difference from the evaluation using the MRR. In both cases, only two pairs of the systems that adjoined in the rank by the MRR interchanged their ranks. That suggests that the MRR is considerably stable for measuring system accuracy on Task 1.

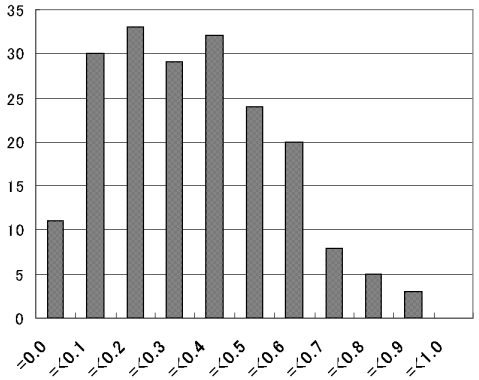

Figure 4. Average over every system's RR of the question in Task 1

Figure 4 is the histogram of the difficulty of the question set of the Task 1. The difficulty of each question is calculated as the average of the reciprocal ranks all the systems achieved to that question. For eleven

questions out of 195 questions (five questions with no answer are excluded), no system could return the correct answer. The easiest question is `QAC1-1099-01`, whose MRR is 0.87 and thirteen systems returned the correct answer in the first rank to this question. The distribution has a smooth curve with one peak, and there are no evidence that the difficulties of the questions are divided into two extremes, that is, too difficult and too easy. The question set used in Task 1 can be concluded to be suitable to evaluate the state of the art of the QA research.

Task 2

Eleven systems participated in the Task 2. The accuracies the participated systems achieved in the mean F-measure (MF) are depicted in Figure 5. The most accurate system achieved 0.36 in the MF. This system always returns a list with one item, and 40% of its answers agree with one of the correct answer items. Another system always returns a list with ten items, and 45% of its answers includes at least one of the correct items, and achieved only 0.09 in the MF. The former strategy is more effective in the current question set, as more than three fourth of the questions have just one correct answer. Other systems seem to determine the number of items included in its answer list dynamically according to a given question. We should examine several measures for Task 2 in order to obtain a measure that reflects our intuition in the goodness on this task.

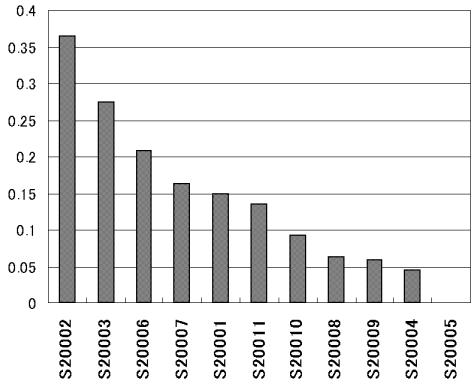

Figure 5. Mean F-measure of participant systems in Task 2

Figure 6 is the histogram of the difficulty of the question set of the Task 2. The difficulty of each question in this case is calculated as the average of the F-measures all the system achieved to that question. Thirty questions out of 200 questions could not answered by any system. The number of such questions is much larger than in Task 1, though the comparison may be meaningless as the measures are different. The easiest question is `QAC1-2136-01`, whose MF is 0.46. Since we used the same question set for both Task 1 and Task 2, we can discuss the characteristics of each task and the relationship between them. We can see some relationships according to the natural expectations. Out of eleven questions that no system answered correctly in Task 1, eight questions were not answered by any system in Task 2 either. Ten easiest questions in Task 1 and Task 2 according to each measure have six overlaps (`QAC1-XYYY-01` where `X=1` or `2`, `YYY=018, 037, 099, 114, 119, 125,` and `136`). Further examination is needed on those relationships.

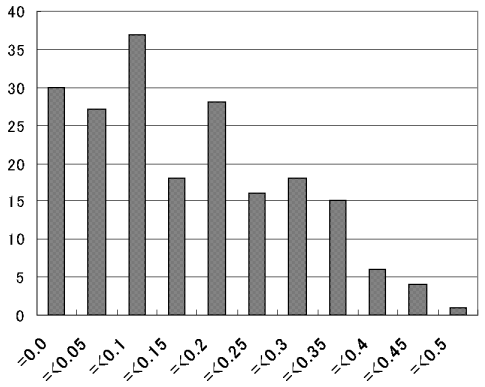

Figure 6. Average over every system's F measure of the question in Task 2

Task 3

Task 3 had only six systems participated and number of questions is just 40. We must be careful to discuss tendencies on this task in this situation. Figure 7 shows accuracies the participated systems achieved in the MF, the same measure as one employed in Task 2. In this task, each problem consists of two successive questions, and the second question, which contains some anaphoric elements, is object to be evaluated. The most accurate system achieved 0.17 in the MF. Fourteen questions out of 40, about one third, could not be answered by any system. We should examine thoroughly the characteristics of this task based on these results and call for more participants.

8.2 Question type and system performance

We have analyzed relationship between the types of questions used for Formal Run evaluation and performance of participant systems in Table 3. "Qnum"

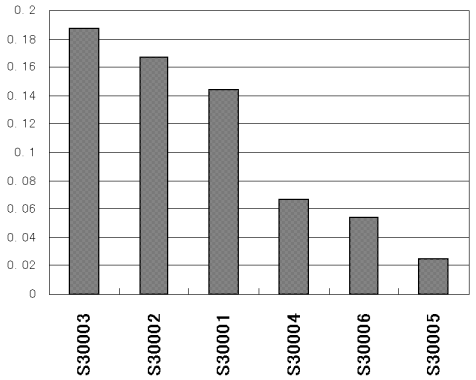

Figure 7. Mean F-measure of participant systems in Task 3

means the number of questions in each question type and "correct" shows the average number of correct system in Task 1 evaluation.

Table 3. System performance analysis for each question type in Task 1

question type	Qnum	correct
artifact name	51	5.4
person name	44	6.8
numeric value	21	5.4
location name	15	6.1
date	14	7.6
country name	13	8.7
company name	12	6.3
other name	12	2.8
organization name	7	5.9
distance	3	7
money	3	7.7
time	3	3
quantity	1	3
percentage	1	6

8.3 System features

It should be emphasized that several architectures or techniques have been tried and employed in the participant systems, though it is difficult to discuss the relations between those attempts and the achieved system performance shown in the previous subsection. For the answer extraction, which extracts answer candidates from retrieved texts or passages, methods using numerical measures are still predominant, in which text is treated as a sequence of words and distance between keywords and answer candidates characterized by a NE tagger plays an important role. Some promising attempts can be found, however, such as those based on matching of syntactic or semantic structures or logical forms. Although meticulously hand-crafted knowledge was still invaluable, machine learning techniques were employed for acquiring several kinds of knowledge of the systems let alone for NE tagging. On the other hand, many systems also use existing tools for their morphological analysis and document retrieval. It can be believed that the infrastructures have been ready for many researchers challenging question answering research. The matter worth special mention is that in addition to system developments many related activities were conducted, which include proposing methods of error analysis, constructing corpus of questions and challenges to speech driven question answering.

9 Conclusion

We have described outline of Question Answering Challenge (QAC1). We have defined three kinds of QA task using two-year newspaper articles and evaluation method for the tasks. We have reported the results of these tasks in terms of statistical results based on MRR and MF and discussed difficulty level of questions of each task from the point of view of average of systems' performance.

We are planning to conduct the second evaluation of QAC as QAC2 at NTCIR Workshop 3 scheduled in May 2004. We will continue analysis of the results from the various aspects and develop better task definition for QAC2.

Acknowledgements

We would like to express our thanks to all the task participants and member of the organizing committee, and also give thanks to staff of NII for their support and having a chance to have this kind of evaluation.

References

[1] J. Fukumoto and T. Kato. An overview of Question and Answering Challenge (QAC) of the next NTCIR workshop. In *Proceedings of the Second NTCIR Workshop Meeting*, pages 375–377, 2001.

[2] J. Fukumoto, T. Kato, and F. Masui. Question and Answering Challenge (QAC-1) : Question answering evaluation at NTCIR workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting Part IV: Question Answering Challenge (QAC-1)*, pages 1–10, 2002.

[3] J. Burger, C. Cardie. et.al. Issues, tasks and program structures to roadmap research in question & answering (q&a), 2001. NIST DUC Vision and Roadmap Documents, http://www-nlpir.nist.gov /projects/duc/roadmapping.html.

[4] E.M. Voorhees and D.M. Tice. Building a question answering test collection. In *Proceedings of SIGIR2000*, pages 200–207, 2000.
[5] Information retrieval and extraction exercise (IREX). http://cs.nyu.edu/cs/projects/proteus/irex/.
[6] Proceedings of 7th Message Understanding Conference (MUC-7), DARPA, 1998.
[7] http://trec.nist.gov/.

Appendix A: Data Formats

The documents used in Dry Run and Formal Run, are Mainichi Newspaper (1998-1999). The document set in the CD-ROM has to be converted using the program, mai2sgml[2], and the output of this conversion is the standard document files for the dry and formal run. Any information not included in this output, such as keywords attached to each article, is considered additional/extra knowledge, which does not included in original newspaper articles.

Format Description

In the following format description, unless specified others, one byte characters are used for all numbers and alphabets. A [xxx] type notation stands for non-terminal symbols, and <CR> represents carriage return.

Question File Format

The Question File consists of lines with the following format.

[QID]: ”[QUESTION]”<CR>

[QID] has a form of [QuestionSetID]-[QuestionNo]-[SubQuestionNo]. [QuestionSetID] consists of four alphanumeric characters. [QuestionNo] and [SubQuestionNo] consists of five and two numeric characters, respectively. [QUESTION] is a series of two byte characters “、” and “。” are used for punctuation marks. “？” is not used.

Examples

QAC0-10001-00: ”大学審議会の会長は誰ですか。”<CR>
QAC0-10002-00: ”タージ・マハーミはどこにありますか。”<CR>
QAC0-10003-00: ”千葉県の県庁淳在地は何市ですか。”<CR>
QAC0-30001-01: ”９８年のアカデミー賞在品賞を受賞した在品は。”<CR>
QAC0-30001-02: ”何という名前の男優が主演したのですか。”<CR>

Answer File Format

The Answer File consists of lines with the following format (so called CSV format).

[QID](, ”[Answer]”, [ArticleID], [HTFlag], [Offset])*<CR> where (...)* is Kleene star, and specifies zero or more occurrences of the enclosed expression.

[QID] is the same as in the question file format above. It must be unique in the file, and ordered identically with in the corresponding question file. It is allowed, however, that some of [QID]s do not list at the file.

[Answer] is the answer to the question, and a series of two byte characters.

[ArticleID] is the identifier of the article or one of the articles used in the process of deriving the answer. The value of the <DOCNO>tag is used for the identifier, which consists of nine numeric characters.

[HTFlag] is “H” or “T”. It will be “H” if the part of article used for deriving the answer and specified in [ArticleID] is the headline, which is the part tagged with <HEADLINE>, “T” if it is the text, which is the part tagged with <TEXT>. This is optional, and when omitted, it should be the empty string , that is, two delimiters, i.e. commas, appear consecutively.

[Offset] is the position of the part used for deriving the answer in the headline or body of text. That position should be represented using number of characters from the beginning of the headline or text. The head of them is represented as zero. A space placed at the beginning of paragraphs is included into characters, while carriage return is not included. This is optional, and when omitted, it should be the empty string , that is, two delimiters, i.e. commas, appear consecutively. “The part of article used for deriving the answer” in the above explanation is typically the portion of the articles where your system extracted the answer from. It does not mean that systems should extract the answer from articles. If your system does not use such extraction for deriving answers, please give us the most relevant position to judge the correctness of your answer. If you can’t specify that anyway, you may omit [HTFlag] and [Offset].

For each question, the quad-gram of ”[Answer]”, [ArticleID], [HTFlag], and [Offset] is repeated more than zero times. In task one, the order of this quad-grams represents the order of the confidence. That is, the most confident

[2]We can obtain this program “mai2sgml” from the URL address, http://lr-www.pi.titech.ac.jp/tsc/tsctools/index-jp.html.

answer candidate should be placed first. The number of candidates is up to five in the dry run. In task two and three, as the answer is a set, the elements of the answer are listed in an arbitrary order.

In the answer file, the line beginning with "#" is a comment. You may include any information, such as a support or context of your answer, as comments.

Examples

The following is an example of the answer to the question:

QAC0-10001-00: "大学審議会の会長は誰ですか。"

It is postulated that the answer is derived using the article shown the below. Three answer candidates are listed. Although all the [ArticleID] are identical in this example, it is not the case in general.

QAC0-10001-00, "石設忠雄", 980701002, T, 24, "町村信孝", 980701002, T, 42, "大学審提言", 980701002, H, 0<CR>

<DOC>
<DOCNO>980701002</DOCNO>
<SECTION> １面 </SECTION>
<AE> 無 </AE>
<WORDS>713</WORDS>
<HEADLINE> 大学審提言「勉強させみ大学」に－－卒業へ評価厳格化 </HEADLINE>
<TEXT>
「２１世紀の大学像」を検討していみ大学審議会（石設忠雄会長）は３０日、中間まとめを町村信孝文相に提出した。単位まとめ取み防止や厳格な成績評価で「勉強しなくて（以下略）
</TEXT>
</DOC>

Appendix B: Usage of Scoring Tool

This tool can be used on command line. As an input argument, a filename of the user's system output should be given. In addition, some expressions such as below options can be used.

–answer / **-a filename** Specifies the filename of the correct answer set. The character code in the file needs to be same as the one used in the user's system output.

–help / **-h** Shows help.

–version / **-v** Shows version of the program.

–task / **-t number** Selects tasks. A number, 1, 2, or 3 follows this option. 1, 2, 3 are for TASK1, TASK2, TASK3, respectively.

–extract / **-e number** Shows the inner data. A number, 1, 2, 3, or 4, follows this option.

The number "1" shows information on each question including, question ID, the total number of answers, the number of different answers, answer number, answer, article ID.

ex.1

```
QAC1-1084-01    15    9
1    法隆          990131022
2    冬大          980521199
2    冬大          981126218
3    薬師          981126218
3    薬師          981230150
4    興福          981126218
⋮    ⋮                  ⋮
```

The number "2" shows the answers that the user's system output including, question ID, the number of answers, answer number, answer, article ID.

ex.2

```
QAC1-1084-01    6
0    法隆          990131022
0    法隆          990131023
1    冬救タミー    980521199
2    冬大          981126218
3    バーミヤ      981126218
4    薬師          981230150
```

The number "3" shows information in detail on correct answers that user's system output. The information includes the correct answers and the answer numbers that correspond to the answer numbers in the correct answer set. The symbol '-' that precedes the answer number means that the answer is correct, but the article ID might not be correct.

ex.3

```
法隆   |   1
法隆   |   -1
冬大   |   2
薬師   |   3
 :     :    :
```

The number "4" shows the score given to each question in Task2. The option is valid for only Task2. Question ID, the number of correct answers, the number of answers that user's system output, the number of correct answers that user's system output and F-measure score for each answers that user's system output.

ex.4

```
QAC1-2146-01:   1   5   1   0.333333
QAC1-2147-01:   1   1   1   1.000000
QAC1-2148-01:   2   5   0   0.000000
QAC1-2149-01:   3   1   1   0.500000
      :         :   :   :       :
```

5 shows the result of answer checking. Question ID, question, list of correct answer and whole answers that user's system output. The correct answer that user's system output are marked with asterisk. The option is valid for only Task1.

ex.5

```
QAC1-1046-01　“奈良の世界遺産にはどのような
ものがありますか”
CORRECT ANSWER:　薬師寺 冬大寺 法隆寺
 平醤宮跡 興福寺 春日大社 春日山原始林
 唐招提寺 元興寺
法隆寺 ＊
冬救タミー
冬大寺 ＊
バーミヤン
薬師寺 ＊
```