Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access

Teruko Mitamura* Hideki Shima* Tetsuya Sakai** Noriko Kando#
Tatsunori Mori† Koichi Takeda‡ Chin-Yew Lin** Ruihua Song**
Chuan-Jie Lin+ Cheng-Wei Lee@
*Carnegie Mellon University **Microsoft Research Asia #National Institute of Informatics
†Yokohama National University ‡IBM Research - Tokyo
+National Taiwan Ocean University @Academia Sinica
teruko@cs.cmu.edu

ABSTRACT

This paper presents an overview of the ACLIA (Advanced Cross-Lingual Information Access) task cluster at NTCIR-8. The task overview includes: a definition of and motivation for the evaluation; a description of the complex and factoid question types evaluated; the document sources and exchange formats selected and/or defined; the official metrics used in evaluating participant runs; the tools and process used to develop the official evaluation topics; summary data regarding the runs submitted; and the results of evaluating the submitted runs with the official metrics.

Keywords Evaluation Metrics, Automatic Evaluation, Evaluation Tools

1. INTRODUCTION

Current research in QA is moving beyond factoid questions, so there is significant motivation to evaluate more complex questions in order to move the research forward. In the Advanced Cross-Lingual Information Access (ACLIA) task at NTCIR-7, we have evaluated cross-lingual and monolingual QA on complex questions (i.e. events, biographies/definitions, and relationships). Our goal in ACLIA for NTCIR-8 is to develop effective CCLQA[1] evaluations for complex questions as well as factoid questions. We will evaluate end-to-end systems and conduct module-based evaluations for question type analysis, document retrieval and answer extraction.

As a central problem in question answering evaluation, the lack of standardization has been pointed out [1], which makes it difficult to compare systems under a certain condition. In NLP research, system design is moving away from monolithic, black box architectures and more towards modular architectural approaches that include an algorithm-independent formulation of the system's data structures and data flows, so that multiple algorithms implementing a particular function can be evaluated on the same task. Following this analogy, the ACLIA data flow includes a pre-defined schema for representing the inputs and outputs of the document retrieval step, as illustrated in Figure 1. This novel standardization effort made it possible to evaluate cross-lingual information retrieval (CLIR) task called IR4QA (Information Retrieval for Question Answering) in a context of a closely related QA task. During the evaluation, the question text and QA system question analysis results were provided as input to the IR4QA task, which produced retrieval results that were subsequently fed back into the end-to-end QA systems. The modular design and XML interchange format supported by the ACLIA architecture make it possible to perform such embedded evaluations in a straightforward manner. More details regarding the XML interchange schemes and so on can be found on the ACLIA wiki [6].

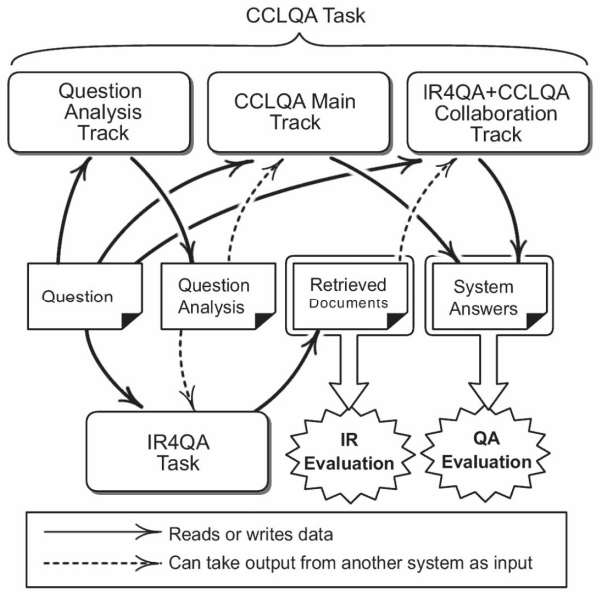

Figure 1. Data flow in ACLIA task cluster showing how interchangeable data model made inter-system and inter-task collaboration possible.

The modular design of this evaluation data flow is motivated by the following goals: a) to make it possible for organizations to contribute component algorithms to an evaluation, even if they cannot field an end-to-end system; b) to make it possible to conduct evaluations on a per-module basis, in order to target metrics and error analysis on important bottlenecks in the end-to-end system; and c) to determine which combination of algorithms works best by combining the results from various modules built by different teams. In order to

[1] Although the task name CCLQA stands for Complex Cross-lingual Question Answering, the task includes evaluation of simple (in terms of information need) factoid questions, and monolingual questions.

evaluate many different combinations of systems effectively, human evaluation must be complemented by development of automatic evaluation metrics that correlate well with human judgment. Therefore, we have developed mechanisms to evaluate the CCLQA results using automatic evaluation methods.

The CCLQA task included topics developed initially for the evaluation of monolingual systems, and then translated to English. Since translation is considered an important challenge in CCLQA, monolingual QA results on the same topics and corpora provide an important baseline for comparison.

The remainder of this paper is structured as follows. Section 2 describes the specific tasks, topics and corpora that were used in the ACLIA evaluation, along with a list of the data interchange formats that were specified to support the exchange of system inputs and outputs. Section 3 provides the details of the CCLQA task definition, and Section 4 provides a detailed explanation of the evaluation metrics and measures (including both human measurement and automatic evaluation techniques). Section 5 provides a brief overview of the support tools that were built to support topic creation and nugget evaluation. Section 6 presents the evaluation results for the participant runs that were submitted. The paper concludes in Section 7.

2. ACLIA TASKS & RUNS

ACLIA consisted of two tasks named IR4QA and CCLQA, which evaluated Information Retrieval and Question Answering, respectively. The same set of topics was shared between the two tasks. For details on the IR4QA task evaluation, we refer the reader to the IR4QA overview paper [2]. This paper focuses on the CCLQA task, which evaluated QA systems on both complex and factoid questions.

For both tasks, we evaluated monolingual and cross-lingual topics. Topic languages included English (EN), Simplified Chinese (CS), Traditional Chinese (CT), and Japanese (JA); target (corpus) languages included CS, CT and JA. An active participant (i.e. a participant who submitted at least one run; hereafter called simply a participant) submitted a system result for at least one of six pairings of source language (i.e. language for questions) and target language (i.e. language for documents), listed below:

- Cross-lingual: EN-CS, EN-CT, EN-JA
- Monolingual: CS-CS, CT-CT, JA-JA

For each track, a participant submitted up to three runs. For each run, we evaluated the top 30 system responses for each question. All official runs were evaluated by independent assessors.

2.1 Evaluation Topics

We focused on the evaluation of five types of complex questions: DEFINITION, BIOGRAPHY, RELATIONSHIP, EVENT, and WHY; four types of factoid questions: PERSON, ORGANIZATION, LOCATION, and DATE. Examples are shown below.

Table 1. Example Questions

Type	#	Example Question	Related Past NTCIR Task
DEFINITION	10	What is the Human Genome Project?	ACLIA
BIOGRAPHY	10	Who is Howard Dean?	ACLIA
RELATIONSHIP	20	What is the relationship between Saddam Hussein and Jacques Chirac?	ACLIA
EVENT	20	What are the major conflicts between India and China on border issues?	ACLIA
WHY	20	Why doesn't U.S. ratify the Kyoto Protocol?	QAC-4
PERSON	5	Who is the Finland's first woman president?	QAC 1-3, CLQA 1,2
ORGANIZATION	5	What is the name of the company that produced the first Fairtrade coffee?	QAC 1-3, CLQA 1,2
LOCATION	5	What is the name of the river that separates North Korea from China?	QAC 1-3, CLQA 1,2
DATE	5	When did Queen Victoria die?	QAC 1-3, CLQA 1,2

A topic developer created a topic by first generating a question and a narrative-style information need in the target language, which were subsequently translated into English. This approach supported a comparison between monolingual and cross-lingual QA using the same set of topics and corpora. A group of volunteers from the participant group created a set of pilot training topics so that details of the task definitions could be refined and finalized.

For the formal evaluation, an independent third-party organization created 100 topics for each target language. Some of the topics are *shared topics* which contain a question originally created for another target language.

2.2 Corpus

The target corpus consists of digital newswire articles (see Table 2). We select newswire articles in the same time span (ranging from 2002 through 2005) in order to support the evaluation of shared topics.

Table2. Corpora used in ACLIA2.

Language	Corpus Name	Time Span	# document
CS	Xinhua	2002-2005	308,845
CT	UDN	2002-2005	1,663,517
JA	Mainichi	2002-2005	377,941

2.3 Input/Output Format

In order to combine a CLIR module with a CLQA system for module-based evaluation, we defined five types of XML

schema to support exchange of results among participants and submission of results to be evaluated:

- **Topic format:** The organizer distributes topics in this format for formal run input to IR4QA and CCLQA systems.
- **Question Analysis format:** CCLQA participants who chose to share Question Analysis results submit their data in this format. IR4QA participants can accept task input in this format.
- **IR4QA submission format:** IR4QA participants submit results in this format.
- **CCLQA submission format:** CCLQA participants submit results in this format.
- **Gold Standard Format:** Organizer distributes CCLQA gold standard data in this format.

For more details regarding each interchange format, see the corresponding examples on the ACLIA wiki [6].

3. CCLQA TASK

Participants in the CCLQA task submitted results for the following four tracks:

- **Question Analysis Track:** Question Analysis results contain key terms and answer types extracted from the input question. These data are submitted by CCLQA participants and released to IR4QA participants.
- **CCLQA Main Track:** For each topic, a system returned a list of system responses (i.e. answers to the question), and human assessors evaluated them. Participants submitted a maximum of three runs for each language pair.
- **IR4QA+CCLQA Collaboration Track:** Using possibly relevant documents retrieved by the IR4QA participants, a CCLQA system generated QA results in the same format used in the main track. Since we encouraged participants to compare multiple IR4QA results, we did not restrict the maximum number of collaboration runs submitted, and used automatic measures to evaluate the results. Participants were able to use the full list of IR4QA results available for each question (up to 1000 documents per topic).

In the CCLQA task, there were six participating teams (see Table 3). The number of submitted runs is shown in Table 4 for the CCLQA main and Question Analysis tracks, and in Table 5 for the IR4QA+CCLQA collaboration tracks.

Table 3. CCLQA Task Participants.

Team Name	Organization
APQA	The University of Tokushima
DLUT	Dalian University of Technology
KECIR	Shenyang Institute of Aeronautical Engineering
LTI	Language Technologies Institute, Carnegie Mellon University
NTOUA	National Taiwan Ocean University
WHUQA	Wuhan University

Table 4. Number of CCLQA runs submitted, followed by number of Question Analysis submissions in parenthesis.

Team Name	CS-CS	EN-CS	CT-CT	EN-CT	JA-JA	EN-JA
APQA	2 (1)					
DLUT	1					
KECIR	2 (1)					
LTI					3 (1)	3 (1)
NTOUA			3	3		
WHUQA	1 (1)	1 (2)				
Total by lang pair	6 (3)	1 (2)	3	3	3 (1)	3 (1)
Total by target lang	7 (5)		6		6 (2)	

Table 5. Number of IR4QA+CCLQA Collaboration runs submitted

Team Name	CS-CS	EN-CS	CT-CT	EN-CT	JA-JA	EN-JA
APQA						
DLUT						
KECIR						
LTI					5	4
NTOUA			3	9		
WHUQA						
Total by lang pair	0	0	3	9	5	4
Total by target lang	0		12		9	

3.1 Answer Key Creation

In order to build an answer key for evaluation, third party assessors created a set of weighted nuggets for each topic. A "nugget" is defined as the minimum unit of correct information that satisfies the information need. In the rest of this section, we will describe steps taken to create the answer key data.

3.1.1 Answer-bearing Sentence Extraction

A nugget creator searches for documents that may satisfy the information need, using a search engine. During this process, a developer tries different queries that are not necessarily based on the key terms in the question text. Whenever a developer finds an answer-bearing sentence or paragraph, it is saved with the corresponding document ID.

3.1.2 Nugget Extraction

A nugget creator extracts nuggets from a set of answer-bearing sentences. In some cases, multiple answer-bearing sentences map to one nugget because they represent the same meaning, even though the surface text is different. In other cases, multiple nuggets are extracted from a single answer-bearing sentence.

3.1.3 Nugget Voting

After nuggets are extracted, we wish to assign weights ranging from 0 to 1 to each nugget in order to model its importance in answering the information need.

In earlier TREC evaluations, assessors made binary decisions as to whether a nugget is *vital* (contains information

to satisfy the information need) or *ok*. More recently, TREC/TAC adopts a pyramid nugget evaluation inspired by research in text summarization. In a pyramid evaluation, multiple assessors make a *vital/ok* decision for each nugget, and weights are assigned according to the proportion of *vital* scores assigned [3].

We adapted the pyramid nugget voting method for the ACLIA evaluation. For each language, there were multiple independent assessors who voted on answer nuggets. Inter-assessor agreement was measured via Fleiss' Kappa statistic, as shown in Table 6. The observed measurements suggest that it would be risky to rely on votes from a single assessor; in this evaluation, each nugget was assessed by all assessors.

Table 6. Inter-assessor agreement on vital/non-vital judgments on nuggets, measured by Fleiss' Kappa.

Language	# of nuggets	# of assessors	Kappa (Inter-assessor agreement)
CS	423	3	0.458
CT	1375	9	0.179
JA	919	3	0.163

We also compared the total number of nuggets and their average character length and weight over the set of topics (see 0).

Table 7. Macro-average nugget statistics over topics.

Lang	Type	Avg #	Avg Char Length	Avg Weight
CS	DEF	4.1	24.9	0.50
	BIO	6.5	18.3	0.34
	REL	4.7	30.6	0.42
	EVE	6.0	24.1	0.44
	WHY	4.0	29.8	0.48
	PER	**1.0**	3.4	0.93
	LOC	1.0	9.6	0.87
	ORG	1.2	9.6	0.97
	DAT	1.4	4.7	0.87
CT	DEF	25.8	19.3	0.78
	BIO	13.4	14.4	0.67
	REL	9.7	21.3	0.80
	EVE	26.8	29.5	0.82
	WHY	11.9	19.9	0.76
	PER	1.2	8.4	0.92
	LOC	1.2	6.3	0.94
	ORG	1.4	14.0	0.91
	DAT	2.2	20.9	0.87
JA	DEF	11.1	14.1	0.74
	BIO	21.0	13.4	0.57
	REL	8.4	17.4	0.73
	EVE	13.0	14.4	0.73
	WHY	7.3	16.0	0.68
	PER	1.0	9.2	1.00
	LOC	1.4	3.6	0.96
	ORG	1.4	9.0	1.00
	DAT	1.4	5.7	1.00

4. EVALUATION METRICS

In this section, we present the evaluation framework used in ACLIA, which is based on weighted nuggets. To avoid the potential ambiguity of the word "answer" (i.e. as in "system answer" and "correct answer"), we use the term *system responses* or *SRs* to denote the output from a CCLQA system given a topic. The term *gold standard* denotes a piece of information that satisfies the information need.

Both human-in-the-loop evaluation and automatic evaluation were conducted using the same topics and metrics. The primary difference is in the step where nuggets in system responses are matched with gold standard nuggets. During human assessment, this step is performed manually by human assessors, who judge whether each system response nugget matches a gold standard nugget. In automatic evaluation, this decision is made automatically. The subsections that follow, we detail the differences between these two styles of evaluation.

4.1 Human-in-the-loop Evaluation Metrics

In CCLQA, we evaluate how good a QA system is at returning answers that satisfy information needs on average, given a set of natural language questions.

In an earlier related task, NTCIR-6 QAC-4 [10], each system response was assigned to one of four levels of correctness (i.e. A, B, C, D); in practice, it was difficult for assessors to reliably assign system responses to four different levels of correctness. For CCLQA, we adopt the nugget pyramid evaluation method [3] for evaluating CCLQA results, which requires only that human assessors make a binary decision whether a system response matches a gold standard *vital* or *ok* nugget. This method was used in the TREC 2005 QA track for evaluating definition questions, and in the TREC 2006-2007 QA tracks for evaluating "other" questions.

A set of system responses to a question will be assigned an *F*-score calculated as shown in Figure 2. We evaluate each submitted run by calculating the macro-average *F*-score over all questions in the formal run dataset.

In the TREC evaluations, a character allowance parameter *C* is set to 100 non-whitespace characters for English [4]. In the NTCIR-7 ACLIA, we adjusted the *C* value for each language based on the character length statistics of the nuggets in the formal run dataset. In the NTCIR-8 ACLIA, we varied *C* value depending on answer type (see Table 7, column Avg Char Length). In this way, we can accommodate questions with longer answers (e.g. C=24.9 in CS DEF) and questions with short answers (e.g. C=3.4 in CS PER) in the same evaluation framework.

Let

r	sum of weights over matched nuggets
R	sum of weights over all nuggets
a_{HUMAN}	# of nuggets matched in SRs by human
L	total character-length of SRs
C	character allowance per match
$allowance$	$a_{HUMAN} \times C$

Then

$$recall = \frac{r}{R}$$

$$precision = \begin{cases} 1 & \text{if } L < allowance \\ \dfrac{allowance}{L} & \text{otherwise} \end{cases}$$

$$F(\beta) = \frac{(\beta^2+1) \times precision \times recall}{\beta^2 \times precision + recall}$$

Figure 2. Official per-topic F-score definition based on nugget pyramid method.

Note that precision is an approximation, imposing a simple length penalty on the SR. This is due to Voorhees' observation that "nugget precision is much more difficult to compute since there is no effective way of enumerating all the concepts in a response" [5]. The precision is a length-based approximation with a value of 1 as long as the total system response length per question is less than the allowance, i.e. C times the number of nuggets defined for a topic. If the total length exceeds the allowance, the score is penalized. Therefore, although there is no limit on the number of SRs submitted for a question, a long list of SRs harms the final F score.

The $F(\beta=3)$ or simply *F3* score has emphasizes recall over precision, with the β value of 3 indicating that recall is weighted three times as much as precision. Historically, a β of 5 was suggested by a pilot study on definitional QA evaluation [4]. In the more recent TREC QA tasks, the value has been to 3. Figure 3 visualizes the distribution of *F3* scores versus recall and precision.

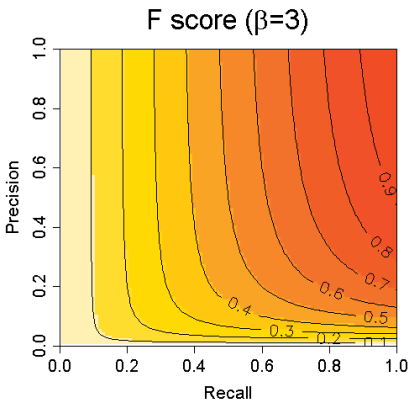

Figure 3. *F3* score distribution parameterized by recall and precision.

As an example calculation of an *F3* score, consider a question with 5 gold standard answer nuggets assigned weights {1.0, 0.4, 0.2, 0.5, 0.7}. In response to the question, a system returns a list of SRs which is 200 characters in total. A human evaluator finds a conceptual match between the 2nd nugget and one of SRs, and between the 5th nugget and one of SRs. Then,

$$recall = \frac{0.4+0.7}{1.0+0.4+0.2+0.5+0.7} = 0.39$$

$$precision = \frac{2 \times 24}{200} = 0.24$$

$$F(\beta=3) = \frac{10 \times 0.24 \times 0.39}{9 \times 0.24 + 0.39} = 0.37$$

The evaluation result for this particular question is therefore 0.37.

4.2 Automatic Evaluation Metrics

ACLIA also utilized automatic evaluation metrics for evaluating the large number of IR4QA+CCLQA Collaboration track runs. Automatic evaluation is also useful during developing, where it provides rapid feedback on algorithmic variations under test. The main goal of research in automatic evaluation is to devise an automatic metric for scoring that correlates well with human judgment. The key technical requirement for automatic evaluation of complex QA is a real-valued matching function that provides a high score to system responses that match a gold standard answer nugget, with a high degree of correlation with human judgments on the same task.

The simplest nugget matching procedure is exact match of the nugget text within the text of the system response. Formally, the assessor a_{HUMAN} in Figure 2 is replaced by $a_{EXACTMATCH}$ as follows:

$$a_{EXACTMATCH} = \sum_{n \in Nuggets} \max_{s \in SRs} \mathrm{I}_{EXACTMATCH}(n,s) \quad (1)$$

$$\mathrm{I}_{EXACTMATCH}(n,s) = \begin{cases} 1 : s \text{ contains } n \text{ in surface text level} \\ 0 : \text{otherwise} \end{cases} \quad (2)$$

Although exact string match (or matching with simple regular expressions) works well for automatic evaluation of factoid QA, this model does not work well for complex QA, since nuggets are not exact texts extracted from the corpus text; the matching between nuggets and system responses requires a degree of understanding that cannot be approximated by a string or regular expression match for all acceptable system responses, even for a single corpus.

For the evaluation of complex questions in the TREC QA track, Lin and Demner-Fushman [8] devised an automatic evaluation metric called POURPRE by replacing a_{HUMAN} with an automatically generated value based on nugget recall:

$$a_{SOFTMATCH} = \sum_{n \in Nuggets} \max_{s \in SRs} \text{NuggetRecall}_{\text{token}}(n, s) \quad (3)$$

$$\text{NuggetRecall}_{\text{token}}(n, s) = \frac{|\text{tokenize}(n) \cap \text{tokenize}(s)|}{|\text{tokenize}(n)|} \quad (4)$$

Since the TREC target language was English, the evaluation procedure simply tokenized answer texts into individual words as the smallest units of meaning for token matching. In contrast, the ACLIA evaluation metric tokenized Japanese and Chinese texts into character unigrams. We did not extract word-based unigrams since automatic segmentation of CS, CT and JA texts is non-trivial; these languages lack white space and there are no general rules for comprehensive word segmentation. Since a single character in these languages can bear a distinct unit of meaning, we chose to segment texts into character unigrams, a strategy that has been followed for other NLP tasks in Asian languages (e.g. Named Entity Recognition [9]).

One of disadvantages of POUPRE is that it gives a partial score to a system response if it has at least one common token with any one of the nuggets. To avoid over-estimating the score via aggregation of many such partial scores, we devised a novel metric by mapping the POURPRE soft match score values into binary values:

$$a_{BINARIZED} = \sum_{n \in Nuggets} \max_{s \in SRs} \text{I}_\theta(n, s) \quad (5)$$

$$\text{I}_\theta(n, s) = \begin{cases} 1 : \text{NuggetRecall}_{\text{token}}(n, s) > \theta \\ 0 : \text{otherwise} \end{cases} \quad (6)$$

We set the threshold θ to be somewhere in between no match and an exact match, i.e. 0.5, and we used this BINARIZED metric as our automatic evaluation metric for ACLIA.

5. EVALUTION TOOLS

To support the creation of test and evaluation topics, as well as the sharing of system and module I/O using XML interchange formats, we developed SEPIA (Standard Evaluation Package for Information Access) [2] which is an improved version the EPAN system used for NTCIR-7. The SEPIA toolkit contains a web interface, a set of utilities and a backend database for persistent storage of evaluation topics, gold standard nuggets, submitted runs, and evaluation results for training and formal run datasets.

5.1 Topic Creation Tools

The SEPIA topic creation tools consist of interfaces for topic development, nugget extraction and nugget voting using the pyramid method. These three activities are described in the subsections that follow.

5.1.1 Topic Development

Figure 4 shows the topic development interface. The left side is the topic creation form, and the right side is an interface to the Lemur/Indri search engine [7], which is used by the topic developer to search for documents relevant to each topic. Topic developers follow these steps:

1. If the developer wishes to modify an existing topic, they can select a topic title from a pull-down list. Topics marked [x] are completed topics. If the developer wishes to start creating a new topic, they can type in the corresponding data and click the "Add" button.
2. Once the developer has created a topic, then they can provide additional information related to the topic: an associated question, a question type, a scenario describing the information need, and a memo containing any extra notes about the topic.
3. In order to search for documents relevant to the topic being created, the developer may directly enter an Indri query, or enter key terms and use the "Generate Query" button to generate an Indri query automatically. When the use is satisfied with the query, it is sent to the Indri retrieval engine.
4. A ranked list of retrieved documents is displayed. The developer can click on a rank number to browse the corresponding full document. When the developer selects a passage which satisfies the information need, the corresponding information is automatically copied into the "Answer Text" and "Doc ID" fields in the Answer data section. The characteristics of the answer-bearing sentences extracted during the ACLIA evaluation are summarized in Section 3.1.1.

5.1.2 Nugget Extraction from Answer Text

Figure 5 shows the nugget extraction interface, which is used to extract nuggets from answer-bearing sentences. (See details in Section 3.1.2)

The user selects a topic title from a list of previously completed topics in the Topic Development task. The user examines the topic data for the selected topic and the answer texts for the selected topic. The users type in the corresponding answer nugget and click "Add" to save the update.

5.1.3 Nugget Voting for Pyramid Method

Figure 6 shows the nugget voting interface, which is used to identify vital nuggets from among the set of nuggets extracted using the nugget extraction tool. (See details in Section 3.1.3).

The user first selects a topic title from a list of previously completed titles in the Topic Development task. The user examines the topic data for the selected topic, and toggles the check boxes next to nuggets which they judge to be *vital*.

[2] http://sourceforge.net/projects/opensepia

5.2 Download and Submission

The SEPIA is used by each participant to upload their submission file for each run submitted. The SEPIA is also used to download intermediate results submitted by other participants, as part of an embedded evaluation, For example, ACLIA participants were able to download the results from Question Analysis and IR4QA in order to conduct an embedded CLIR evaluation.

5.3 Evaluation

The SEPIA provides interfaces for supporting the core human-in-the-loop part of evaluation: relevance judgment for IR4QA and nugget matching for CCLQA. In each task, items to be evaluated belong to a pool created by aggregating the system responses from all systems, based on run priority. For the three runs submitted by each team in each ACLIA task, we created three pools of system responses. For the CCLQA task, the first pool (corresponding to run 1) was evaluated by independent third-party assessors hired by NII. The second and third pools (corresponding to runs 2 and 3) were evaluated by volunteers including members of the participant teams. Details of the CCLQA results are provided in Section 6.1. For the embedded IR4QA collaboration track, the system responses were evaluated automatically; details are provided in Section 6.2.

6. EVALUATION RESULTS

In this section, we will present evaluation results for the CCLQA main track and IR4QA collaboration track.

6.1 CCLQA Main Track

The human evaluation results for CCLQA are shown for each language pair in Tables 8 through 13. Assessors evaluated all (up to 30) system responses per run per question. Unfortunately, an intermittent but serious bug in the evaluation system for CCLQA caused some system responses to be truncated prematurely. Since the bug was detected quite close to the publication deadline for the proceedings, we were unable to correct the problem and re-evaluate the affected systems in time for this paper. A preliminary analysis indicates that the corrected results may be a few percentage points lower in some cases. We plan to repeat the human evaluation for the problematic system responses and release the official results as soon as possible. On a positive note, we were able to fix the truncation problem and re-run the automatic evaluation in time for this publication. The results are shown in Section 6.1.2.

6.1.1 Official Runs (Preliminary Results)

Table 8. EN-CS official human evaluation.

EN-CS Runs	ALL
WHUQA-EN-CS-01-T	0.1841

Table 9. CS-CS official human evaluation.

CS-CS Runs	ALL
APQA-CS-CS-01-T	0.1388
APQA-CS-CS-01-D	0.1422
DLUT-CS-CS-01-T	0.1954
KECIR-CS-CS-01-T	0.3450
KECIR-CS-CS-02-T	0.3354
WHUQA-CS-CS-01-T	0.2566

Table 10. EN-CT official human evaluation.

EN-CT Runs	ALL
NTOUA-EN-CT-01-T	0.1096
NTOUA-EN-CT-02-T	0.1362
NTOUA-EN-CT-03-T	0.1159

Table 11. CT-CT official human evaluation.

CT-CT Runs	ALL
NTOUA-CT-CT-01-T	0.1815
NTOUA-CT-CT-02-T	0.1988
NTOUA-CT-CT-03-T	0.1461

Table 12. EN-JA official human evaluation.

EN-JA Runs	ALL
LTI-EN-JA-01-T	0.1074
LTI-EN-JA-02-T	0.1130
LTI-EN-JA-03-T	0.1045

Table 13. JA-JA official human evaluation.

JA-JA Runs	ALL
LTI-JA-JA-01-T	0.1069
LTI-JA-JA-02-T	0.1443
LTI-JA-JA-03-T	0.1438

6.1.2 Automatic Evaluation results

Table 14. EN-CS automatic evaluation.

EN-CS Runs	ALL
WHUQA-EN-CS-01-T	0.2895

Table 15. CS-CS automatic evaluation.

CS-CS Runs	ALL
APQA-CS-CS-01-T	0.3117
APQA-CS-CS-01-D	0.2668
DLUT-CS-CS-01-T	0.3080
KECIR-CS-CS-01-T	0.4614
KECIR-CS-CS-02-T	0.4502
WHUQA-CS-CS-01-T	0.3513

Table 16. EN-CT automatic evaluation.

EN-CT Runs	ALL
NTOUA-EN-CT-01-T	0.1813
NTOUA-EN-CT-02-T	0.2035
NTOUA-EN-CT-03-T	0.1717

Table 17. CT-CT automatic evaluation.

CT-CT Runs	ALL
NTOUA-CT-CT-01-T	0.2823
NTOUA-CT-CT-02-T	0.2579
NTOUA-CT-CT-03-T	0.1885

Table 18. EN-JA automatic evaluation.

EN-JA Runs	ALL
LTI-EN-JA-01-T	0.2107
LTI-EN-JA-02-T	0.2029
LTI-EN-JA-03-T	0.2040

Table 19. JA-JA automatic evaluation.

JA-JA Runs	ALL
LTI-JA-JA-01-T	0.2024
LTI-JA-JA-02-T	0.2259
LTI-JA-JA-03-T	0.2252

6.2 IR4QA+CCLQA Collaboration Track

Tables 20 through 23 show the results from the IR4QA+CCLQA Collaboration track (see Section 3), evaluated using the automatic evaluation method described in Section 4.2.

Table 20. EN-JA Collaboration Track: F3 score based on automatic evaluation.

		CCLQA
		LTI
IR4QA	CYUT-EN-JA-01-T	0.3169
	CYUT-EN-JA-02-T	0.3252
	CYUT-EN-JA-03-D	0.3358
	CYUT-EN-JA-04-DN	0.3349

Table 21. JA-JA Collaboration Track: F3 score based on automatic evaluation.

		CCLQA
		LTI
IR4QA	BRKLY-JA-JA-01-DN	0.2934
	BRKLY-JA-JA-02-T	0.2686
	BRKLY-JA-JA-03-DN	0.2074
	BRKLY-JA-JA-04-DN	0.3000
	BRKLY-JA-JA-05-T	0.2746

Table 22. EN-CT Collaboration Track: F3 score based on automatic evaluation.

		CCLQA		
		CYUT1	CYUT2	CYUT3
IR4QA	CYUT-EN-CT-02-T	0.1520	0.1679	0.1608
	KDEG-EN-CT-02-DN	0.1726	0.1790	0.1654
	QUTIS-EN-CT-04-T	0.1287	0.1479	0.1466

Table 23. CT-CT Collaboration Track: F3 score based on automatic evaluation.

		CCLQA		
		CYUT1	CYUT2	CYUT3
IR4QA	KDEG-CT-CT-05-T	0.2973	0.2797	0.2649

7. CONCLUSION

This paper presented an overview of the ACLIA (Advanced Cross-Lingual Information Access) task cluster at NTCIR-8, with a specific focus on the CCLQA evaluations. We described the official metrics used in evaluating participant runs; the tools and process used to develop the official evaluation topics; summary data regarding the runs submitted; and the preliminary results of evaluating the submitted runs with the official metric. We also utilized automatic evaluation metrics for both CCLQA evaluation and the IR4QA+CCLQA collaboration tasks.

ACKNOWLEDGEMENTS

We greatly appreciate the efforts of all the NTCIR-8 ACLIA participants, organizers and advisors.

REFERENCES

[1] Lita, L.V. Instance-Based Question Answering. *Doctoral Dissertation*, Computer Science Department, Carnegie Mellon University. 2006.

[2] Sakai, T., H. Shima, N. Kando, R. Song, C.-J. Lin, T. Mitamura, M. Sugimoto. Overview of NTCIR-8 ACLIA IR4QA, *Proceedings of NTCIR-8*, 2010.

[3] Lin, J., and D. Demner-Fushman. Will pyramids built of nuggets topple over?, *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. 2006.

[4] Voorhees, E. M. Overview of the TREC 2003 Question Answering Track, *Proceedings of TREC 2003*, 2004.

[5] Voorhees, E. M. Overview of the TREC 2004 Question Answering Track, *Proceedings of TREC 2004*, 2005.

[6] The ACLIA wiki, http://aclia.lti.cs.cmu.edu/ntcir8

[7] Indri search engine: http://www.lemurproject.org

[8] Lin, J., and D. Demner-Fushman. Methods for Automatically Evaluating Answers to Complex Questions, *Information Retrieval*, 9(5):565-587, 2006.

[9] Asahara, M., Y. Matsumoto. Japanese Named Entity Extraction with Redundant Morphological Analysis, Proceedings of NAACL/HLT 2003, 2003.

[10] Fukumoto, J., T. Kato, F. Masui, T. Mori. An Overview of the 4th Question Answering Challenge (QAC-4) at NTCIR Workshop 6, *Proceedings of NTCIR-6*, 2007.

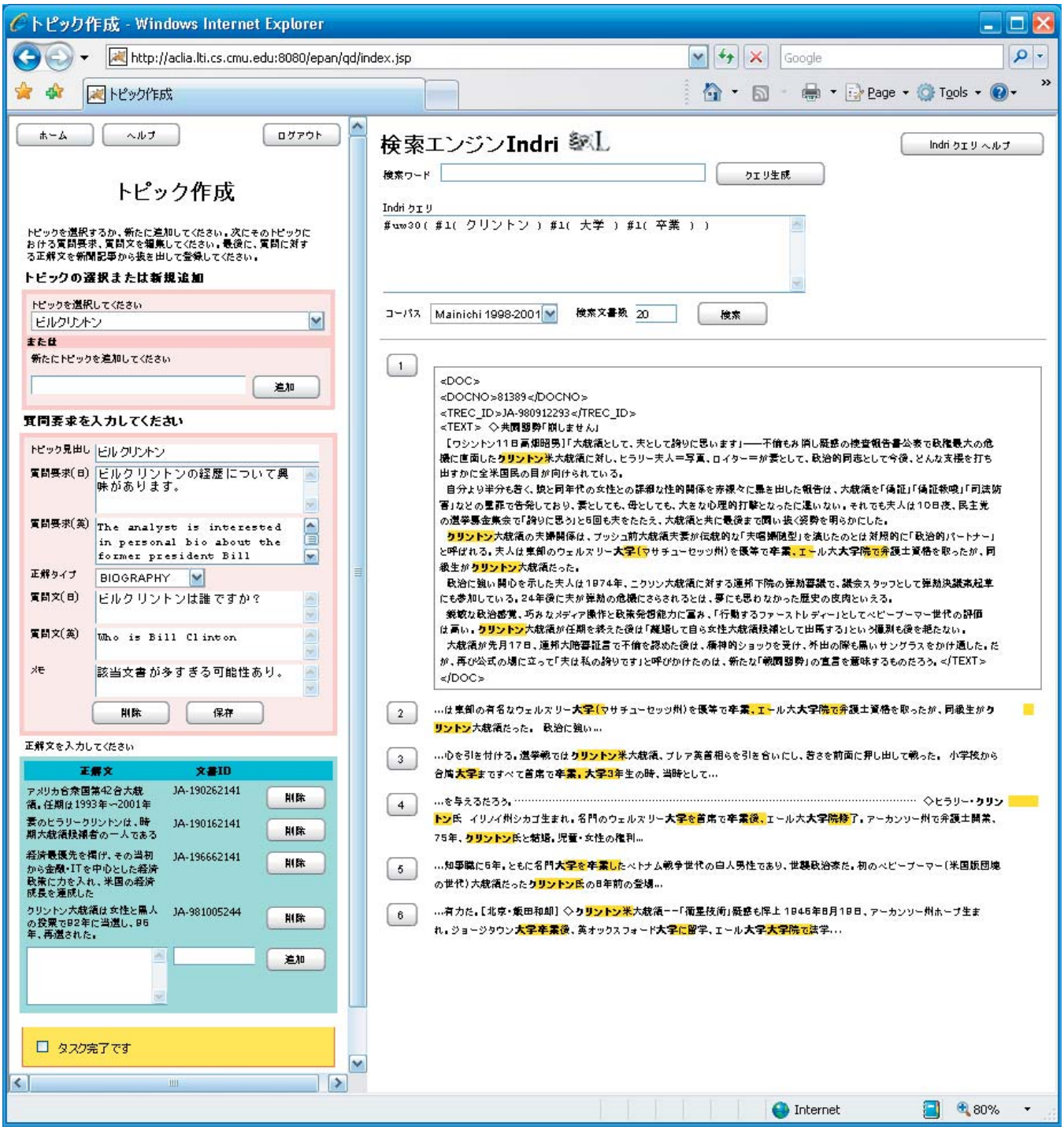

Figure 4. SEPIA topic development interface.

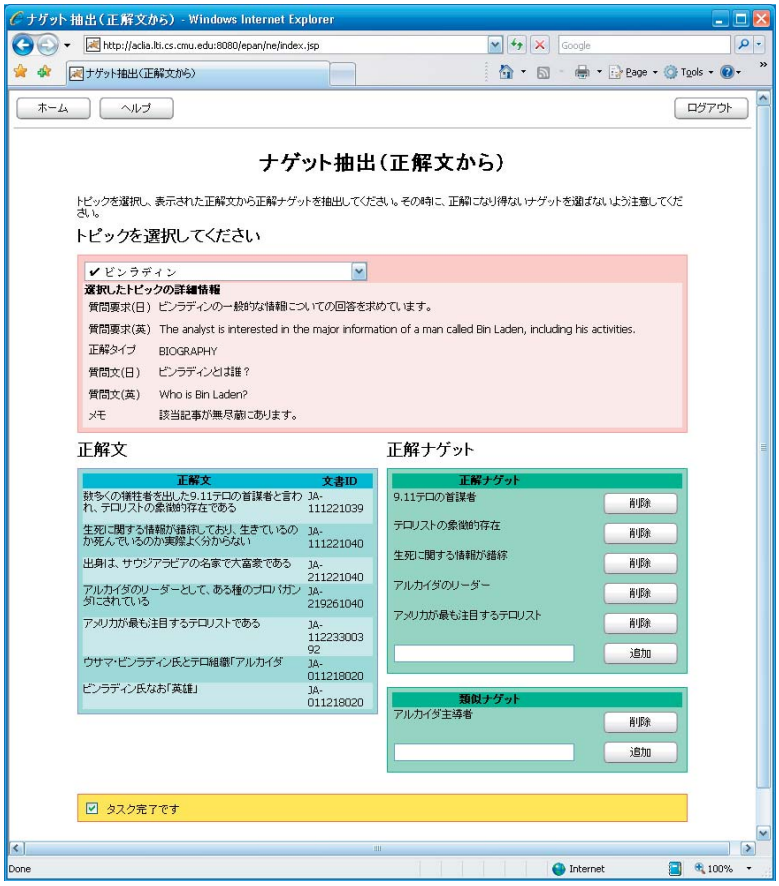

Figure 5. SEPIA nugget extraction interface.

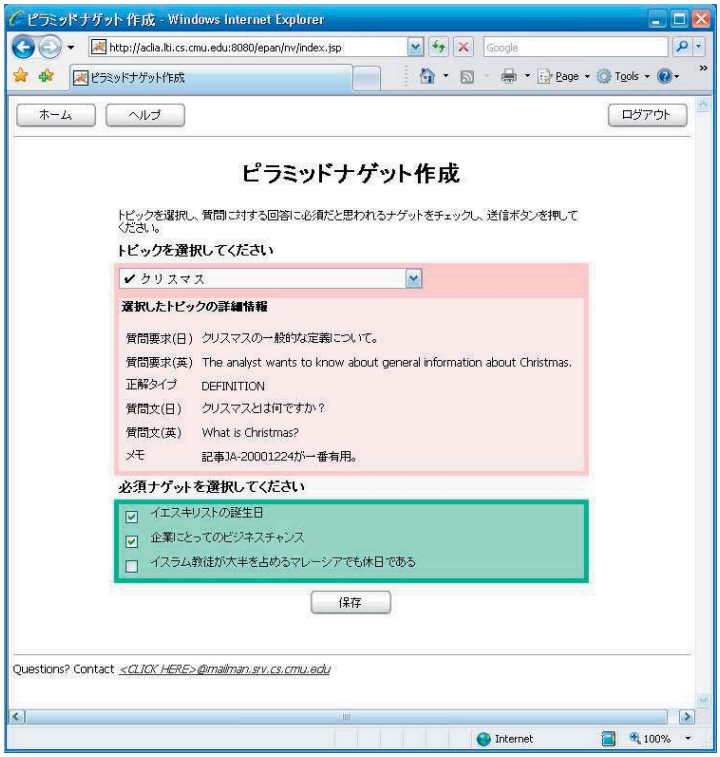

Figure 6. SEPIA nugget voting interface.