Modular Approach to Error Analysis and Evaluation for Multilingual Question Answering

Hideki Shima Mengqiu Wang Frank Lin Teruko Mitamura
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{hideki, mengqiu, frank+, teruko}@cs.cmu.edu

1 Introduction

The JAVELIN system is a modular, extensible and language-independent architecture for building question-answering systems [3]. We have been working to extend the original English version of JAVELIN for cross-language question answering in Chinese and Japanese. Recently, we participated in the NTCIR5 CLQA1[1] evaluation. Out of 13 groups participating in the CLQA1 task, we are the only group to submit formal runs for both the English-to-Chinese (EC) and the English-to-Japanese (EC) subtasks. After analyzing the observed performance of each module on the evaluation data, we created gold standard data (perfect input) for each module in order to determine upper bounds on module performance. This modular approach allows us to compare the performance of the two systems (EC and EJ) on a per-module basis.

2 Javelin Architecture

The JAVELIN system is composed of five main modules: Question Analyzer (QA), Translation Module (TM), Retrieval Strategist (RS), Information eXtractor (IX) and Answer Generator (AG). Inputs to the system are processed by these modules in the order listed above. The QA module is responsible for parsing the input question, choosing the appropriate answer type, and producing a set of keywords. The TM module translates the keywords into task-specific languages. The RS module is responsible for finding relevant documents which might contain answers to the question, using translated keywords produced by the TM. The IX module extracts answers from the relevant documents. The AG module normalizes the answers and ranks them in order of correctness. The overall architecture is shown in Figure 1.

[1]Cross-Lingual Question Answering (CLQA) Task in the Fifth NTCIR Workshop (2004-2005)

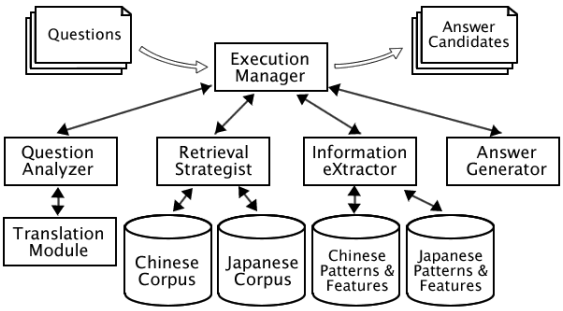

Figure 1. System Architecture

3 Result and Analysis

Our overall submission to the NTCIR CLQA1 task included three submissions for the English-Chinese (EC) subtask, and three submissions for the English-Japanese (EJ) subtask. The three runs for each subtask were carried out using three different IX strategies – FST, Light and Combo IX. FST IX is based on finite-state transducer, Light IX is based on simple distance-based algorithms, and Combo IX is a combination of both. More details on these IX strategies can be found in [2]. 200 input questions were provided for each of the subtasks. For each question, only the top answer candidate that was returned by the system was judged. Correct answers that were not properly supported by the returned document were judged to be unsupported answers.

Formal run results are shown in Table 1. The highest accuracies for EC and EJ tasks are 7.5% and 10.0%, respectively, both achieved by IX Light.

3.1 Module-by-Module Analysis

In order to gain different perspectives on the tasks and our systems' performance, a module-by-module analysis was performed. This analysis was based on gold-standard answer data, which also provides information about the documents that contain the correct answer for each question. We judged the QA mod-

Table 2. Performance from Partially Gold Standard Input

	Gold Standard	QA_{ATYPE}[a]	TM[b]	RS[c]	IX[d] (MRR[e])	Accu[f] (Unsup[g])
EC	None	86.5%	69.3%	30.5%	30.0% (0.130)	7.5% (9.5%)
	TM	86.5%	—	57.5%	50.0% (0.254)	9.5% (20.0%)
	TM+QA_{ATYPE}	—	—	57.5%	50.5% (0.260)	9.5% (20.5%)
	TM+QA_{ATYPE}+RS	—	—	—	63.0% (0.489)	41.0% (43.0%)
EJ	None	93.5%	72.6%	44.5%	31.5% (0.116)	10.0% (12.5%)
	TM	93.5%	—	67.0%	41.5% (0.154)	9.5% (15.0%)
	TM+QA_{ATYPE}	—	—	68.0%	45.0% (0.164)	10.0% (15.5%)
	TM+QA_{ATYPE}+RS	—	—	—	51.5% (0.381)	32.0% (32.5%)

[a] Average presicion of answer-type detection
[b] Average presicion of keyword translation over 200 formal run questions
[c] Average precision of document retrieval. Counted if correct document was ranked between 1st–15th
[d] Average precision of answer extraction. Counted if correct answer was ranked between 1st–100th
[e] The MRR measure of IX performance, calculated by averaging the sum of the reciprocal of each answer's rank
[f] Overall accuracy of the system
[g] Accuracy including unsupported answers

Table 1. Formal Run Performance

	EC		EJ	
	Corr[a]	Unsup[b]	Corr[a]	Unsup[b]
FST	14 (7.0%)	19 (9.5%)	17 (9.0%)	20 (10.0%)
LIGHT	15 (7.5%)	19 (9.5%)	20 (10.0%)	25 (13.0%)
COMBO	10 (5.0%)	12 (6.0%)	17 (9.0%)	20 (10.0%)

[a] Corr – Correct answer
[b] Unsup – Unsupported answer

ule by the accuracy of its answer type classification, and the TM module by the accuracy of its keyword translation. For the RS and IX modules, if a correct document or answer is returned, regardless of its ranking, we consider the module to be successful. To separate the effects of errors introduced by earlier modules, we created gold-standard data by manually correcting answer-type and keyword translation errors. We also create "perfect" IX input using the gold-standard document set.

The results are shown in Table 2. Note that because Light IX performed best in the formal run, for both EC and EJ, we will focus our discussion on Light IX in this paper.

3.1.1 QA Performance

The QA module performed well in identifying the answer type in both subtasks. As we can see from the QA_{ATYPE} column in Table 2, the QA achieved 86.5% for the EC subtask and 93.5% for the EJ subtask. An additional analysis of accuracy by answer type is shown in Table 3. Compared to row *TM+QA_{ATYPE}* in Table 2, we can see that further improvement of the answer type accuracy via manual correction did not make a significant difference.

	EC			EJ		
A-type	# of Q	correct	%	# of Q	correct	%
PER	79	64	81%	34	34	100%
LOC	45	44	98%	34	33	97%
ORG	15	12	80%	13	8	62%
ARTI	27	23	85%	21	19	90%
DATE	18	18	100%	25	25	100%
TIME	1	1	100%	14	13	93%
MONEY	5	4	80%	20	18	90%
NUMEX	10	7	70%	31	29	94%
PCNT	0	0	-	8	8	100%
Sum	200	173	86.5%	200	187	93.5%

Table 3. QA Performance by Answer Type

3.1.2 TM Performance

The average precision of translation was 69.3% for the EC subtask and 72.6% for the EJ subtask. By taking advantage of translation by web-mining, we could successfully translate some named entities. After manual correction of keyword translation errors, we immediately gained over 20.0% accuracy in the RS module performance for both the EC and EJ subtasks, as shown in row *TM* in Table 2. This shows that translation errors have a significant negative impact on keyword-based document retrieval.

3.1.3 RS Performance

The RS module achieved an accuracy of 30.5% in the EC subtask and 44.5% in the EJ subtask, as shown in column *RS* in Table 2. To illustrate the difference before and after manual translation of keywords, a CLIR-style analysis of the RS module is provided in Table 4. For all the questions that showed an improved MRR score after manual correction of keyword translation errors, the TM failed to translate 43 and 88 keywords in the EJ and EC subtasks, respectively. Among these

	No man-trans		With man-trans	
Rank	EC	EJ	EC	EJ
1	11	29	44	52
2-5	30	31	38	53
6-9	14	12	20	15
10-15	6	17	13	14
no match	139	111	85	66
Sum	200	200	200	200
MRR	0.12	0.22	0.31	0.37
Success Rate	30.5%	44.5%	57.5%	67.0%

Table 4. RS Evaluation: Number of correct documents by retrieved rank

keywords, 65.0% for the EJ subtask and 43.0% for the EC subtask were classified as proper nouns and phrases by the QA module. Most of the proper nouns are person, location and organization names. We also observed that in the corpus, the majority of the questions were drawn from these three types. This helps to explain the 20.0% accuracy gain achieved from corrected key term translation.

3.1.4 IX Performance

In the formal run data (row *None* in Table 2), we observed big accuracy drops at the RS module and after the IX module for both the EC and EJ subtasks, and bigger accuracy drops at the IX module for the EJ subtask. The drop in RS accuracy is expected, but the difference between Light IX performance in the EC and EJ subtasks is surprising. After eliminating errors carried over from earlier modules, the IX in the EC and the EJ subtasks show a performance difference of 11.5%(63.0%-51.5%); see row $TM+QA_{ATYPE}+RS$.

The Light IX used the same algorithm in the EC and EJ subtasks, but with different distance measure functions and different parameter settings. The IX in the EC subtask achieved a higher MRR score in all cases, and better accuracy in most cases (except in the formal run). But the EC system had worse overall accuracy than the EJ system, except in the $TM+QA_{ATYPE}+RS$ case. We cannot conclude at this point which Light IX setting is more effective, because other factors such as corpus tagging precision differences are involved. In general however, we found the Light IX in the EC system to be more accurate and produced more answer candidates.

Because the answer validation function was not yet implemented in the AG module to filter out noise, the overall accuracy of the EC and EJ systems is much lower than the accuracy of the IX module in both cases. We can see the degradation caused by the noise in IX output by examining the $TM+QA_{ATYPE}$ row and $TM+QA_{ATYPE}+RS$ row in the EC part of Table 2. The accuracy of the IX differs only by 12.5%(63.0%-50.5%), but this measure does not take into account noise in other answer candidates. The effect of the noise is delayed until the output of the AG module, where a 31.5%(41.0%-9.5%) difference in overall answer accuracies and a 22.5%(43.0%-20.5%) difference including unsupported answers are seen.

As the performance of the RS increased after manual correction of keyword translation errors, the IX module showed a similar increase in performance of 20.0%(50.0%-30.0%) in the EC subtask and 10.0%(41.5%-31.5%) in the EJ subtask. But as we increase the accuracy of RS from 57.5% in EC and 67.0% in EJ to 100.0%, by manually creating "perfect" RS output, the performance of the IX module did not increase as much. The upper bound on IX performance was 63.0% for the EC subtask and 51.5% for the EJ subtask.

4 Issues and Proposed Solutions

From the modular analysis, we observed low performance of the IX module on numerical and temporal questions. What we also noticed is that for these types of questions, subtype information such as 'year', 'percentage' is very informative and could be used to improve IX performance. Table 1 showed Combo IX did not work as effectively as we expected. It is difficult to decide how much recall should be sacrificed for accuracy when an IX module is used in combination with others. The JAVELIN system for English incorporates a Planner module which can select among the set of available IX modules at run-time [1]. It is our future work to adapt it for use in CLQA.

5 Conclusion

Our analysis of per-module performance from gold-standard input shows that the QA module and the RS module are already performing fairly well, but there is still room in the IX module and the AG module for future improvement. Also, we found that keyword translation accuracy greatly affects overall performance on the CLQA task.

References

[1] L. S. Hiyakumoto. Planning in the JAVELIN QA System. *Carnegie Mellon Computer Science Technical Report CMU-CS-04-132*, 2004.

[2] F. Lin, H. Shima, M. Wang, and T. Mitamura. CMU JAVELIN System for NTCIR5 CLQA1. *To appear in Proceedings of the Fifth NTCIR Workshop*, 2005.

[3] E. Nyberg, T. Mitamura, J. Callan, J. Carbonell, R. Frederking, K. Collins-Thompson, L. Hiyakumoto, Y. Huang, C. Huttenhower, S. Judy, J. Ko, A. Kupsc, L. V. Lita, V. Pedro, D. Svoboda, and B. V. Durme. The JAVELIN Question-Answering System at TREC 2003: A Multi-Strategy Approach with Dynamic Planning. *In Proceedings of TREC 12*, November 2003.