Question Answering Performance on Table Data

Xing Wei, Bruce Croft, David Pinto
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01002
{xwei,croft,pinto}@cs.umass.edu

ABSTRACT

Question answering (QA) on table data is a challenging information retrieval task. This paper describes a QA system for tables created with both machine learning and heuristic table extraction methods. Errors were analyzed in order to improve the system using government statistical data. We also apply these improvements on another type of table data set and show the experimental results.

1. INTRODUCTION

Tables are an important source for question answering (QA) systems. They provide a visual way to link metadata (headers, titles) with cell data (question answers). Linking that data and metadata, especially from text tables formatted by humans, is a difficult task. The table itself must be identified within the text document and data and header lines separated. Within the set of header rows, an algorithm must recognize the difference between titles and column headers and determine the span of each individual column header. Finally, row headers are identified. With this information in place, the data is combined, creating an answer passage for each table cell.

There are a number of approaches to these decisions. Previously, a heuristic system [2] and a machine learning system [3] using conditional random fields (CRF) decided on labels for individual lines of text files. This paper describes the answer retrieval experiments with CRF extractor and improvements. The answer retrieval results with heuristics extractor were presented as a reference.

2. METHODS

We used two methods to extract our data, one based on heuristics and one based on CRFs. Then we built the database of cell documents using Lemur [1].

During answer retrieval, first, the system retrieves the documents for a query from the database using Lemur's structured query language. Secondly, passages in each document are ranked and the best scoring passage in each document is evaluated as a potential answer for the question.

The passages were ranked using language modeling techniques.

$$P(A \mid Q) \propto \prod_{i=1}^{n} P(q_i \mid A) \qquad (1)$$

To evaluate the system, we have two training sets from www.FedStats.gov. One is the original data set and the other is the enlarged one. We also have one training set of Wall Street Journal (WSJ) data. In the QA retrieval tests there are 50 questions for FedStats tables and 55 questions for WSJ tables.

3. ANSWER RETRIEVAL FROM TABLES

The results on FedStats data are in Table 1.

Table 1. MRR for QA retrieval baselines

	MRR at	Original CRF	Heuristic
No stemming/ stopping	1	0.14	0.14
	5	0.178	0.171
	100	0.194	0.187
With stemming/ stopping	1	0.1	0.14
	5	0.149	0.168
	100	0.171	0.187

From the results above we can see that heuristic extraction works almost the same or even better than CRF extraction on QA retrieval. An analysis of the errors indicated that the problem is related to low recall on header lines by the CRF. Table header lines are very important. They are the link that connects table cells with a query.

In order to improve the answer retrieval results, we tried a two-step improvement. The first step trained a new CRF extraction trained with more features and data; the second step added the algorithm that is more tolerant of non-table lines in the headers.

Table 2 shows the improvements in MRR in percentage terms. Results are shown for both the improvements over the heuristic extraction method and extraction using the original CRF.

Table 2. Percentage Improvements

	MRR at	Vs. Heuristic		Vs. Original CRF	
		Step 1	Final	Step 1	Final
No stemming/ stopping	1	14%	43%	14%	43%
	5	9%	40%	5%	35%
	100	6%	37%	2%	32%
With stemming/ stopping	1	14%	43%	60%	100%
	5	15%	48%	30%	66%
	100	12%	42%	23%	55%

To confirm these improvements, we test the system on the Wall Street Journal data, which contains many simpler forms text tables than FedStats. The results are in Table 3.

Table 3. MRR for QA retrieval from WSJ data

	MRR at	Original CRF	Heuristic	Improved CRF
No stemming/ stopping	1	0.055	0.145	0.109
	5	0.092	0.211	0.135
	100	0.102	0.219	0.146
With stemming/ stopping	1	0.018	0.127	0.091
	5	0.061	0.188	0.137
	100	0.075	0.200	0.146

The QA retrieval results from the simpler WSJ tables extracted by CRFs are also improved considerably by the same ideas that worked for the more complex tables of FedStats. But the results are still below the heuristic extractor. The WSJ tables are editorially different than the FedStats tables. Training the CRFs with FedStats data shows that this CRF may not generalize to all table styles. Increasing the training set to include tables from various sources, such as the WSJ, may make the CRF extraction more robust. We trained a new CRF on WSJ data and achieved much better QA performance in Table 4. QA retrieval performance was tested with 53 of the original 55 questions, as two of them happened come from tables in the training set.

4. DISCUSSION

In this paper we attempted to improve the performance of question answering on table data with the CRF extraction model. The CRF extractor finally achieved better QA results than the heuristic extractor on FedStats data did.

Table 4. MRR for QA retrieval, CRF trained on WSJ

	MRR at	Heuristic	CRF
No stemming/ stopping	1	0.145	0.377
	5	0.211	0.482
	100	0.219	0.491
With stemming/ stopping	1	0.127	0.377
	5	0.188	0.465
	100	0.200	0.471

The experiments with the WSJ dataset indicate that the heuristic approach is stable across many types of tables. But CRF is more promising. Its QA retrieval results are improved by the methods we applied in these experiments. Inclusion of more varied tables, such as in the WSJ database, should lead to increased performance.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the National Science Foundation under NSF grant #EIA-9983215 and in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] Ogilvie, P. and Callan, J. Experiments using the Lemur toolkit. In *Proceedings of the 2001 Text REtrieval Conference (TREC 2001),* 103-108, National Institute of Standards and Technology, special publication 500-250, 2002.

[2] Pinto, D., Croft, W. B., Branstein, M., Coleman, R., King, M., Li, W. and Wei, X. Quasm: A system for question answering using semi-structured data. In *Proceedings of the JCDL 2002 Joint Conference on Digital Libraries*, 46–55, 2002.

[3] Pinto, D., McCallum, A., Wei, X. and Croft, W. B., Table Extraction Using Conditional Random Fields, In *Proceedings of SIGIR '03 Conference,* 235-242, 2003.