Rule Extraction for Automatic Question Answering Based on Structural Clustering

Shen Song, Yu-N Cheah, Enya Kong Tang, Bali Ranaivo-Malançon

School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia

Summary
Automatic question answering (QA) is playing an increasingly important role in intelligent answer searching. Many approaches have been employed for retrieving answers to natural language questions with rule-based approach being one of them. Traditionally, rules for automatic QA have been generated manually which may be time consuming and limited in scope. To address this issue, we present a proposed automatic rule extraction approach to generate rules for QA from training data via structural clustering.

Key words:
Automatic question answering, rule extraction, structural clustering.

1. Introduction

Users of information technology (IT) have long faced the problem of retrieving relevant information from various resources such as text documents and the Internet. This is largely due to the sheer information overload faced by IT users. Search engines have been proven useful in addressing many keyword-related search initiatives. However, their effectiveness lies in the skill of the users to construct the right queries.

To further facilitate the search for information and to improve the user interface, automatic question answering (QA) approaches have been developed to allow questions to be posed in natural language [1]. These QA systems avoid the need for users to formulate structured queries in order to retrieve a particular piece of information. Another added advantage is that QA systems also have the potential to respond to a user's query in natural language.

Traditionally, rule-based approaches have been employed in QA systems for the matching mechanism in view that it was simple, efficient and effective. Generally, the rule-based approach for QA involves the manual generation of rules. One reason for this is that QA rules aim to generalise the way natural language questions are answered. Due to the complexities of natural languages (such as grammar, styles, etc.) it was deemed appropriate for QA rules to be authored manually by humans.

Manuscript received March 5, 2008
Manuscript revised March 20, 2008

Manual rule generation, however, may result in redundant and less-than-optimum rules. Manual generation of rules is also time-consuming. Therefore, to address these issues, we present an approach to automatically extract rules for QA by sentence structural clustering [2]. In our proposed approach, we utilise the Synchronous Structured String-Tree Correspondence (S-SSTC) [3] and Example-Based Machine Translation (EBMT) [4] structure indexing tools to help analyse data, i.e. the question and answer sentences.

2. State-of-the-Art QA Systems

QA is a specialised field in information retrieval. Besides rule-based approaches, various other techniques have been employed to facilitate the retrieval of relevant answers to users' queries. Some QA systems are introduced here.

2.1 WEBCOOP

WEBCOOP provides intelligent cooperative responses to web queries by the integration of knowledge representation and the use of advanced reasoning procedures. However, the "cooperativity" only focuses on atomic and enumerative responses. The system offers two modes for querying a web page either via keywords or in natural language. Constraints can be relaxed to prevent returning a wrong answer. The system utilises a knowledge extractor and a robust question parser to select and examine the proposed answers. According to the cooperative rules, the WEBCOOP inference engine will determine the matching answers and organise them for output [5].

2.2 AskMSR

This system (see Figure 1) depends on data redundancy rather than on the linguistic resources. Even in question reformulation, the rewrites are only simple string-based manipulations. The n-gram technique is used to retrieve possible answer. After filtering the n-grams, the system applies an answer tiling algorithm to combine similar

answers and assemble longer answers from overlapping smaller answer fragments [6].

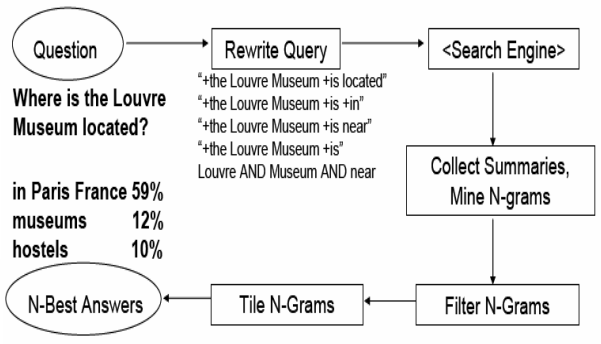

Figure 1: AskMSR system architecture [6]

In this system, strategies are also explored for predicting when the QA system is likely to give an incorrect answer.

2.3 Quarc

Semantic and lexical heuristics are employed in Quarc [7] to capture answers for a given question in a reading comprehension setting. In view that different "wh" questions (i.e. who, what, when, where, and why) require different types of answers, a separate set of rules are generated manually. A score is assigned to candidate sentences which fulfil the conditions set in the rules. An example for "what" rules are shown in Figure 2. In Quarc, the entire sentence which is deemed to answer the question will be presented to the user.

```
1. Score(S) += WordMatch(Q,S)
2. If contains(Q,MONTH) and
      contains(S,{today, yesterday,
      tomorrow, last night})
   Then Score(S) += clue
3. If contains(Q,kind) and
      contains (S, {call,from})
   Then Score(S) += good_clue
4. If contains(Q,name) and
      contains( S, {name,call,known})
   Then Score += slam_dunk
5. If contains(Q,name+PP) and
      contains(S,PROPER_NOUN) and
      contains(PROPER_NOUN,head(PP))
   Then Score(S) += slam_dunk
```

Figure 2: "what" rules [7]

2.4 A Comparative Overview

Generally, a wide variety of techniques have been employed to address QA. As mentioned earlier, WEBCOOP utilises an inference engine with advanced reasoning (with question refinement capabilities). AskMSR employs decision trees and relies on data redundancy. Quarc on the other hand is a lexical and semantic-heuristic rule-based system.

The rule-based approach has been used widely in various research fields. While not discounting the effectiveness of other approaches mentioned earlier, we choose to revisit the simple, effective and efficient approach of rules to see how we can improve it further. As mentioned earlier, the major drawback about current rule-based approaches in QA is that the generation of such rules is not automated. While automated rule extraction is not new in other domains for knowledge discovery, we believe that the automatic generation or rather the extraction of QA rules is quite novel and present its own challenges.

3. Automatic Rule Extraction for Question Answering

Generally, natural language questions constructed in a particular way are also answered in a particular way. Therefore, our methodology aims to induce rules to match questions to answers by structural clustering of question and answer sentences.

Our methodology consists of three phases (see Figure 3):

- compilation of question-answer pairs
- analysis of question-answer pairs
- rule extraction via clustering.

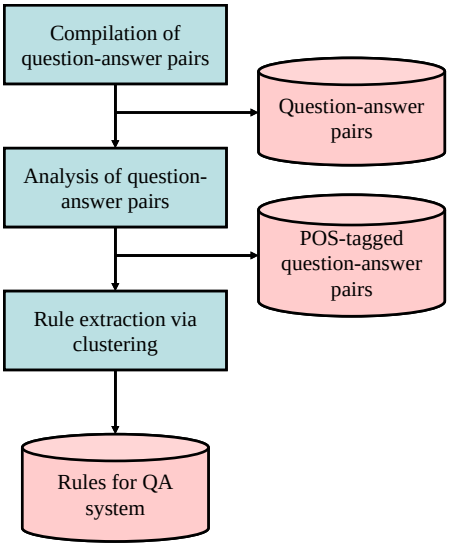

Figure 3: Methodology for QA rule extraction

3.1 Compilation of question-answer pairs

A large number of question-answer pairs is needed to support our methodology. We decided on the CBC Reading Comprehension Corpus by the Canadian Broadcasting Corporation's CBC 4 Kids website and made available by MITRE Corporation. This corpus contains 125 news articles. Each article is accompanied by a list of questions. POS-tagged and answer-tagged versions of the corpus are available. Since the answer-tagged version of the CBC Reading Comprehension Corpus tags the whole sentence that contains the answer to a particular question, and does not answer the question in a direct way, we have found it necessary to reword the answers.

From the corpus, we selected 400 question-answer pairs as our database and these are grouped into "what", "where", "when" and "who" questions. An example of "what" question-answer pairs is as shown in Table 1.

Table 1: "what" question-answer pairs

No	*Question*	*Answer*
1	What is the purpose of an anti-personnel landmine?	The purpose of an anti-personnel landmine is to incapacitate, injure or kill someone.
2	What is the resolution of Canada's Radarsat-2 satellite?	The resolution of Canada's Radarsat-2 satellite is three meters.
3	What does Lloyd say is the message from NATO?	The message from NATO is that military force against civilians is not acceptable.

3.2 Analysis of question-answer pairs

In this phase, two main activities are carried out:

- obtaining the correspondences between the respective questions and answers
- indexing of parts-of-speech (POS) phrases

In the first activity, the question-answer pairs are tagged with POS and the correspondences between the respective questions and answers are obtained. This is achieved using the S-SSTC tool [3]. With this tool, we are also able to visualise the questions and answers in a tree structure. Examples (from Table 1) are shown in Figures 4a, 4b and 4c) respectively. In each of the three question-answer pair's tree structure, the left panel shows the question tree structure while the right panel illustrates the answer tree structure. The tables at the bottom-right corner display the correspondences between the respective question and answer to indicate similar nodes between the question and answer. The back-end processing of the S-SSTC tool produces a tree structure file of the question-answer sentences. We then use this file as input to the EBMT indexing tool for POS phrase indexing.

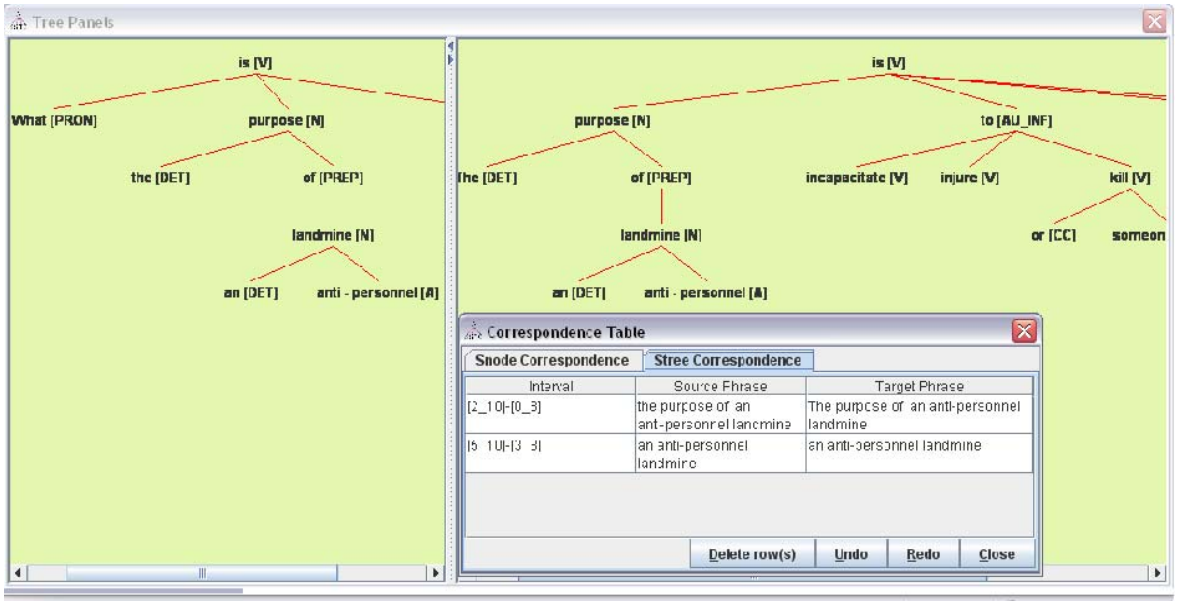

Figure 4a: S-SSTC output for question-answer pair no. 1

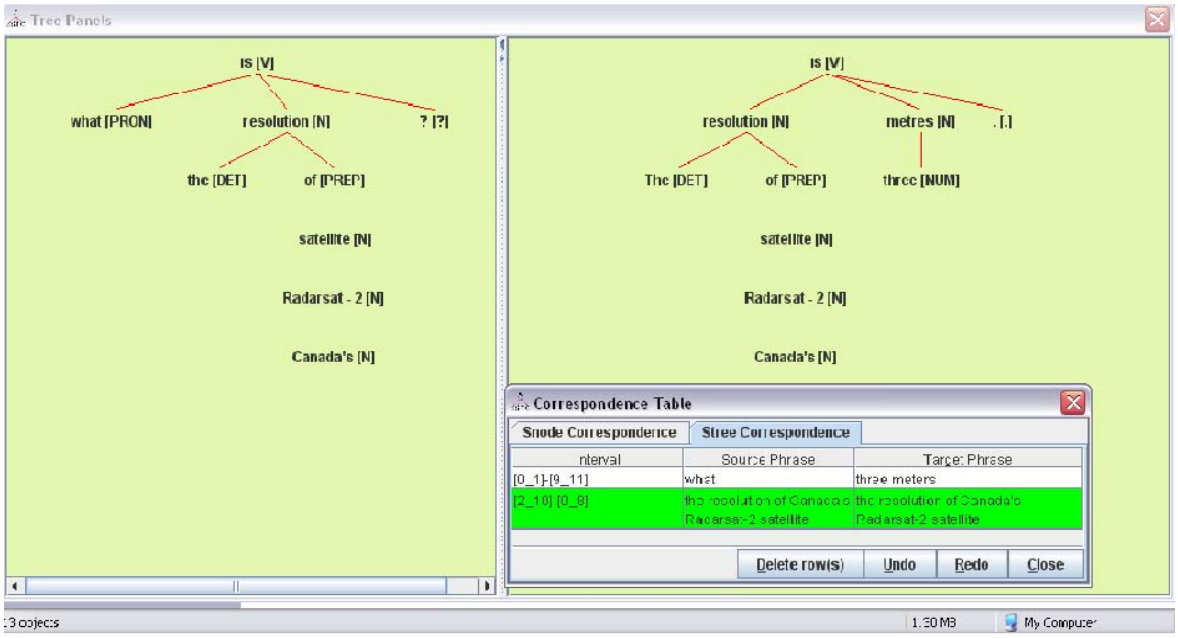

Figure 4b: S-SSTC output for question-answer pair no. 2

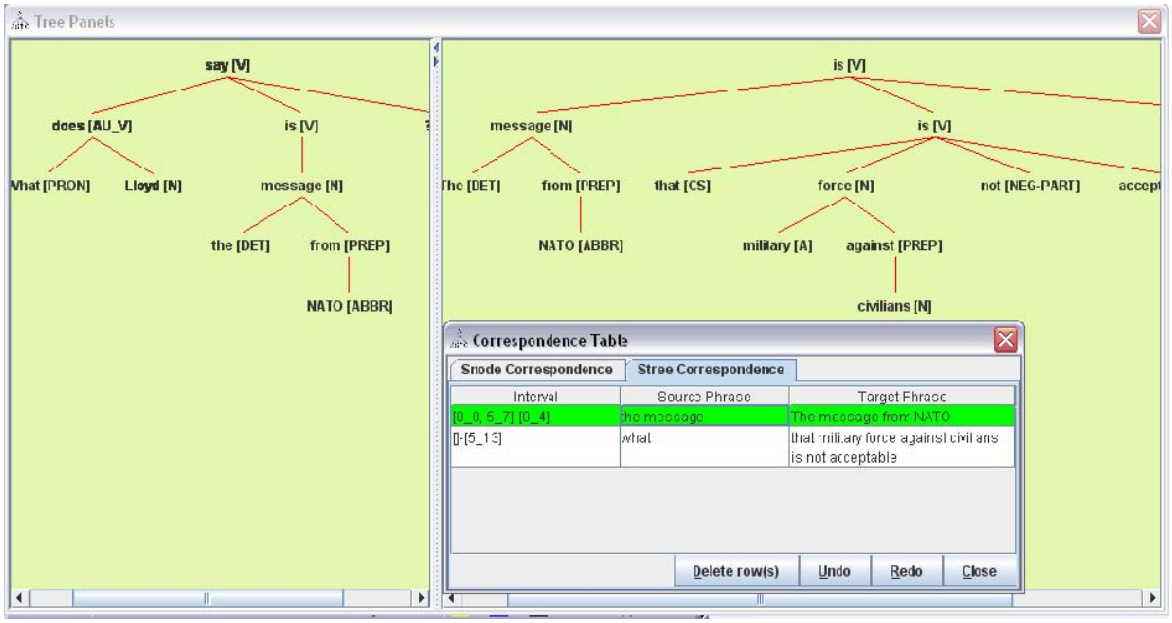

Figure 4c: S-SSTC output for question-answer pair no. 3

In the second activity using the indexing tool [4], only the question tree structure part is indexed. The indexed examples are as shown in Table 2. The EBMT indexing tool produces an index according to POS phrases. The last column of the table indicates the corresponding question number.

Table 2: Questions indexed according to POS phrases

POS Phrase No.	*POS Phrase*	*Tree Layer*	*Question No.*
1	VVP	1	3
2	VBZ	0	1, 2
3	JJ	1	1
4	DT NN	1	1, 2, 3
5	IN	2	1, 2, 3
6	VVZ	2	3
7	NP	1	2, 3
8	POS	0	2

In preparation for the next phase, the original index produced by the EMBT indexing tool (from Table 2) is inverted, thus producing an inverted index according to the questions with their respective POS phrases (see Table 3).

Table 3: Questions indexed according to question numbers

Question No.	*POS Phrases of Sentence*	*POS Phrase No.*
1	WP VBZ DT NN IN DT JJ NP SENT	2 4 5 3 7
2	WP VBZ DT NN IN NP POS NP NN SENT	2 4 5 7 8 7
3	WP VVZ NP VVP VBZ NP IN DT NN SENT	6 7 1 2 7 5 4

3.3 Rule Extraction via Clustering

Using the inverted index (Table 3), the questions will then be clustered based on the similarity distance of POS phrase numbers and their sequence (see Figure 5).

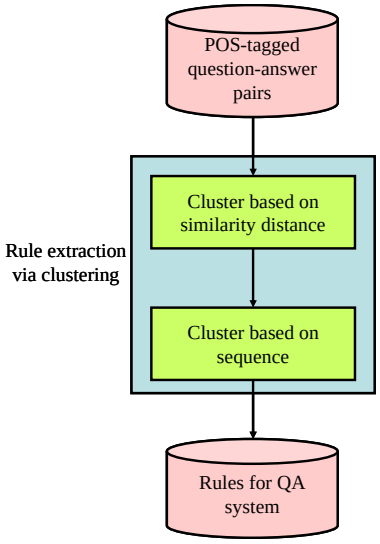

Figure 5: Clustering for QA rule extraction

Three strategies of clustering the question and answer pairs are possible [8]:

- Cluster only the question part
- Cluster only the answer part
- Cluster both the question and answer parts together

In our methodology, we will cluster only the question part and select the best answer structure that corresponds to that question structure based on popularity.

In clustering the questions, firstly, we consider the similarity distances of the POS phrases (represented by the POS phrase numbers). The more similar the POS phrase numbers between two questions, the higher the likelihood that the respective questions will belong to the same cluster. This is indicated by a scoring mechanism that keeps track of the question numbers of other questions that are similar to a particular question. After analysing the similarity distances, the questions will be pre-clustered. As an example, it is obvious that all the three questions have POS phrase 2, 4, 5, 7 (indicated with underlined text in Table 3). The similarity distances, p_n, between them may be indicated as follows:

$$p_{n(Q1,Q2)} = p_{n(Q2,Q3)} = p_{n(Q1,Q3)}$$

After the initial pre-clustering, we then proceed to cluster according to sequence similarity. Sequence similarity focuses on the similar POS phrase sequence in each question. Using the earlier three-question example, we observe that the sequence of the POS phrases in question no. 3 is quite different from the other two, while question nos. 1 and 2 have a similar sequence of their POS phrases.

To combine the above pre-clustering (similarity distance) and sequence clustering steps, the BLEU [9] similarity measurement technique can be simply employed. Based on the original BLEU formula, we present a modified BLEU definition as follows:

$$\text{BLEU} = \exp \sum_{n=1}^{N} w_n \log p_n$$

Where $w = 1 - \frac{sequences}{matches}$ and $p = \frac{matches}{lengths}$. w is the weight for the sequence similarity of the sentence pairs with *sequences* indicating the number of differences in the POS phrase sequence (i.e. if any two sentences have the same POS phrase sequence, then *sequences* = 0). The value of *matches* indicates the number of similar POS phrases in any two sentences, while *lengths* indicate the number of POS phrases present in the longer sentence of any two sentences. The BLEU similarity measurement technique is based on n-grams. For our purpose, we are only examining the sentences word-by-word, and hence, N = 1.

This concludes the clustering phase which, using the example in Table 3, results in 2 clusters, i.e. one cluster containing question nos. 1 and 2, while another containing question no. 3.

Lastly, within each cluster, we then analyse the respective answer parts and choose the top *n* most popular answer structures to pair up with the respective question part. This will be necessary in the event that some individual question structures may be answered using more than one answer structure.

After the POS phrase and sequence clustering, as well as the selection of relevant answer parts, the following rules are extracted (see Table 4):

Table 4: Extracted Rules

Rule	*Question*	*Answer*
1	WP VBZ ( )$_{NP}$ IN ( )$_{NP}$ SENT	( )$_{NP}$ IN ( )$_{NP}$ VBZ TO ( )$_{VV}$ NN SENT
		( )$_{NP}$ IN ( )$_{NP}$ VBZ ( )$_{NP}$ SENT
2	WP VVZ ( )$_{NP+VVP}$ VBZ ( )$_{NP}$ IN ( )$_{NP}$ SENT	( )$_{NP}$ IN ( )$_{NP}$ VBZ ( )$_{THAT\ clause}$ SENT

4. QA Example using the Extracted Rules

As an example of a matching process, let us assume we would like to answer the question, "Q: What is the name of the Portuguese anti-landmine film?". Assume also that a repository of POS-tagged documents (which may contain the answers) is available.

The question will be analysed and converted into the POS tagged form as follows: WP VBZ DT NN IN DT JJ NN NN SENT. Meanwhile, the question will also be analysed in a more general manner, such as combining DT and NN into $(\ )_{NP.}$ Then the final question form will be: WP VBZ $(\ )_{NP}$ IN $(\ )_{NP}$ SENT.

Based on the extracted rules above, we observe that the question belongs to Rule 1, and the corresponding answer structure should be either $(\ )_{NP}$ IN $(\ )_{NP}$ VBZ TO $(\ )_{VV}$ NN SENT or $(\ )_{NP}$ IN $(\ )_{NP}$ VBZ $(\ )_{NP}$ SENT. From this, an attempt will be made to retrieve sentences from the POS-tagged document repository which matches the suggested answer structure from the rule. From these candidate answers, certain individual words will be further examined to determine which of these would answer the question exactly. Therefore, the result will be answers in natural language form.

5. Conclusion and Future Work

In this paper, we introduced an automatic rule extraction methodology via structural clustering. In our approach, we utilised two similarity measurements to perform the clustering: (1) POS phrase number similarity and (2) its sequence similarity in each sentence. Our research efforts are still in progress, particularly in the area of similarity measurement, i.e. with the exploration of n-grams and introducing parameters such as the sentence length. This, we hope, would extend and improve the effectiveness of the BLEU similarity measurement technique. The S-SSTC and EBMT indexing tools employed in this work also proved to be useful. However, we look forward towards making this process more efficient in delivering the desired answers to users.

Acknowledgements

This research is funded by the Malaysian Ministry of Science, Technology and Innovation via its ScienceFund research grant.

References

[1] Hagen, P.R., Manning, H. and Paul, Y., Must Search Stink, *Forrester Research*, June 2000.

[2] Song, S. and Cheah, Y.-N., Extracting Structural Rules for Matching Questions to Answers, *Workshop on Language, Artificial Intelligence and Computer Science for Natural Language Processing Applications (LAICS-NLP)*, Bangkok, Thailand, 2006.

[3] Al-Adhaileh, M.H. and Tang, E.K., Synchronous Structured String-Tree Correspondence (S-SSTC), *IASTED International Conference on Applied Informatics (AI 2002)*, Innsbruck, Austria, 2002.

[4] Al-Adhaileh, M.H. and Tang, E.K., Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema, *Machine Translation Summit VII*, Singapore, 1999, pp. 244-249.

[5] Benamara, F. and Dizier, P.S., WEBCOOP: A Cooperative Question-Answering System on the WEB, *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003*), Budapest, Hungary, 2, 2003, pp. 63-66.

[6] Brill, E., Dumais, S. and Banko, M., An Analysis of the AskMSR Question-Answering System, *Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA, 2002.

[7] E. Riloff and M. Thelen. A rule-based question answering system for reading comprehension tests, *ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, Washington, USA, 2000.

[8] Lin, D. and Pantel, P., Discovery of Inference Rules for Question Answering, *Natural Language Engineering*, 7(4), 2001, pp. 343-360.

[9] Shen, Y., Zaccak, G., Katz, B., Luo, Y. and Uzuner, O., Duplicate Removal for Candidate Answer Sentences, *First CSAIL Student Workshop (CSW)*, Massachusetts, USA, 2006.

Shen Song received her B.Sc. degree in Computer Science from LiaoNing University, China in 2002. She is currently pursuing a master degree at the School of Computer Sciences in Universiti Sains Malaysia, Penang, Malaysia. Her research work is focused on automatic question answering systems.

Yu-N Cheah received his B.Comp.Sc. (Hons.) degree from Universiti Sains Malaysia in 1998, and his Ph.D. degree from the same university in 2002. He is currently a senior lecturer at the School of Computer Sciences, Universiti Sains Malaysia and the head of the Health Informatics Research Group. His research interests include knowledge management, intelligent systems, and health informatics.

Enya Kong Tang is currently an Associate Professor at the School of Computer Sciences, Universiti Sains Malaysia. He obtained his B.Comp.Sc. (Hons.) degree in 1990 and Ph.D. degree in 1994 from Universiti Sains Malaysia. His current research interests are computational linguistics, machine translation and advanced information retrieval techniques.

Bali Ranaivo-Malançon was born in Madagascar. She received a Ph.D. in Natural Language Processing from the National Institute for Oriental Languages and Civilisations (France) in 2001. Since 2002, she has been with the School of Computer Sciences, Universiti Sains Malaysia, where she is currently a Lecturer. Her research interests include the development of linguistic resources through text analysis and corpus processing. Her objective is to provide to NLP applications like question-answering systems the resources that they need to improve their accuracy.