A Student Information System in Restricted Domains

Purushottam Das, Mohd Mursleen, Ankur Singh Bist
Department of Computer Engineering
G.B. Pant University of Agriculture and Technology, Pantnagar
Uttarakhand, India

ABSTRACT
Our approach is tried to achieve high accuracy in Student Information System (SIS) in restricted domain. The restriction leads to the question domains and extract answers. NLP queries get handled and get traversed under the various phases under SIS system for answer retrieval in appropriate defined manner. User query profiling, inference rules and framing are used to infer the answer.
Keywords: NLP queries, retrieval, inference rules, framing, restricted domain

1. INTRODUCTION
During the last decade, automatic question answering has become an interesting research field and resulted in a significant improvement in its performance, which has been largely driven by the TREC (Text REtrieval Conference) SIS Track [1]. Track is able to answer questions correctly 70% of the time [2]. The 70% of accuracy is, of course, high enough to surprise the researchers of this field, but, on the other hand, the accuracy is not enough to satisfy the normal users in the real world, who expect more precise answers. The difficulty of constructing open-domain knowledge base is one reason for the difficulties of open-domain question answering. Since question answering requires understanding of natural language text, the SIS system requires much linguistic and common knowledge for answering correctly. The simplest approach to improve the accuracy of a question answering system might be restricting the domain it covers. By restricting the question domain, the size of knowledge base to build becomes smaller.

This paper describes our restricted domain question answering system for a Student Information System (SIS). One of the roles of the SIS is to retrieve information to any Natural Language Query. The SIS should provide high precision answers; otherwise the users will not trust the entire functions of the SIS, which includes not only the ability of question answering. That means no answer is preferred to a wrong answer and the primary concern in our research is improving the precision of the question answering system and generate same SQL query for similar natural language queries. In this paper, we present a question answering system which is

restricted to answer only to the questions on Student table (Fig. 1). The domain specific hand-coded ontology containing student names and courses is manually built for the question analysis and the inference process. The remainder of the paper is organized as follows. Section 2 describes the overall architecture of the SIS system. Section 3 describes the Student Information System. Section 4 evaluates the system and reports the limitation of the Student Information system. Section 5 compares our system with other Student Information systems. Section 6 concludes with some directions for future work.

2. OVERALL ARCHITECTURE
The overall framework of the student Information system is presented in Figure 2. The SIS performs three-phase processing: First, it analyses natural language questions and translates the questions into Structured Query Language (SQL) statements. Second, the SQL queries are directed to a DBMS to retrieve the answers in the database. Finally, the result from the DBMS is converted to natural language sentences for output. Figure 2 depicts overall processes for the Student Information System. A DBMS is used for managing extracted data. Table shows the details of students having attributes as Names, ID_No, Branch, Courses, Grade and College.

TABLE I
STUDENT TABLE

Names	ID_No	Branch	Courses	Grade	College
John	2201	CSE	DB, DS	A+	COE
Marry	2202	CSE	DB, DS	A++	COE
Jessy	2203	ME	HMT, T	B	COE
Rock	2204	ME	HMT, T	C	COE
Jack	2205	EE	CS, CT	A	COE
Bruce	2206	EE	CS, CT	C	COE

3. STUDENT INFORMATION SYSTEM
The question answering starts from extracting Student information from table. The user request is analyzed with the question analyzer and the appropriate query frame is selected, and then the request is translated into the SQL expression. The SQL query is used to retrieve the correct answer from the database, which stores student information in the Student table. Finally, natural language answer is generated based on the every result extracted from the DBMS.

Flow Control:

Sentences → Keywords Set → Frame → SQL generation → SQL Retrieval form DB → NLP Answer Generation

.

A. Question Analysis

The First, user's request is analysed with the query analyzer as represented in Figure 2. The analyzer extracts several keywords that describing the question, such as name word, course, grade and branch by using a dependency parser and the user question is represented only by these extracted keywords. The named entity tagger is used to identify temporal expressions, names, grade and courses. The tagger consults the domain-dependent ontology for recognizing student names and the domain independent ontology for place names. The ontology for the student names consists of frame concepts, which are similar to Synset in WORDNET [3]. For example, course and areas are in same event concept in the domain ontology for course events, because the questions about areas are usually asking about courses studied.

If the information is not matched in the frames, the question analyzer returns that topic is not relevant to our domain. The inference rules, which are built, based on our observation on various user questions, are domain-independent, because the omission of temporal or spatial information is common not only in weather information question, but also in questions for other domains.

Sentences → Keywords Set

Let's take an example of the query analysis. The following keywords are extracted from the question:

"Write all names of students in engineering."

EVENT : names

PROCEDURE : Write

DOMAIN : engineering

If information is not explicitly mentioned in the question, the question analyzer infers that the information is out of topic. The information can't be retrieved from our restricted domain.

B. Query Frame Decision

Restricting the question domain and information resource, we could restrict the scope of user request. That is, there is a finite number of expected question topics.

Keywords Set → Frame → SQL generation

"Write all names of students in engineering."

Keyword set = [Write, names, engineering]

Frame:

[(Common Keyword) (Synonyms)]

Each expected question topic is defined as a single *query frame*. The following are query frame examples. Frames works as a intermediate representation between NLP sentences and SQL queries. They are used for processing the query for the names, courses and grades. The main frames of our restricted domain are:

[names]

[courses]

[grades]

Each frame has a rule for SQL generation. Frame names has the following SQL generation rule:

SQL generation rule:

[Names]

SELECT Student.names

FROM Student

WHERE BRANCH = "ME" ^ BRANCH= "CSE" ^ BRANCH = "EE" or COLLEGE = "COE"

Names, Courses, Grade and ID No are field names in the database table Student table. The rule generates the SQL statement that means: retrieve the information of frame from the DB table which stores Student information.

(List, Shows, Describe, Illustrate, Choose, Write, Explain)

(Courses, Fields, Area, Subject)

(Grade, Performance, GPA, CGPA, Division)

(names, students)

Here is another example, which is the SQL generation rule for [courses].

SQL Query:

SELECT Student .courses

FROM Student

WHERE BRANCH = "ME" ^
BRANCH = "CSE" ^
BRANCH = "EE" or
COLLEGE = "COE"

Analyzing a question means selecting a query frame in this system. Thus, it is important to select the appropriate query frame for the user request. The selection process is a great influence on the precision of the system, while there is not much likelihood of errors in other processes, such as

generating SQL query from the selected query frame, retrieving DB records, and generating an answer.

The query frame selection is based on the extracted event, procedure and domain. Currently, A hand-coded decision tree-like classifier is used for selecting an appropriate query frame for the extracted keywords. If a question isn't proper for the handling domain, the classifier rejects it. Machine learned classifier is being developed in order to substitute for the hand-coded classifier.

C. SQL Generation

If a query frame is selected for a question, an SQL query statement is generated from the SQL production rule of the frame. The query is sent to the DBMS to acquire the records that match to the query. Following explanation depicts the conversion from a natural language question to its SQL expression.

SQL generation → SQL Retrieval form DB

"Write all names of students in engineering."

↓

EVENT : names

PROCEDURE : Write

DOMAIN : engineering

↓

The frame [names] is selected.

↓

SQL Query:

```
SELECT Student.names

FROM Student

WHERE   BRANCH = "ME" ^
        BRANCH= "CSE" ^
        BRANCH = "EE" or
        COLLEGE = "COE"
```

D. Answer Generation

Based on the result of the DBMS, a natural language answer is generated. A rule based answer generation method is used. Each query frame has an answer generation pattern for the frame.

SQL Retrieval form DB → NLP Answer Generation

For example frame grades has the following generation pattern: [names]

NLP Output:

The **names** of the students are as follows:

$[names]

SQL Query:

```
SELECT Student.names

FROM Student

WHERE   BRANCH = "ME" ^
        BRANCH = "CSE"
        ^BRANCH = "EE" or
        COLLEGE = "COE"
```

↓

DBMS returns:

Names
John
Marry
Jessy
Rock
Jack
Bruce

↓

NLP Output:

The **Names** of students are as follows:

Names
John
Marry
Jessy
Rock
Jack
Bruce

$1 is the field value of the queried result. $1 is the function that gives expression to a natural language expression. [More Sample outputs are given in Appendix]

4. EVALUATION AND LIMITATION

Our domain restricted SIS system is evaluated based on precision and recall, and investigated the limitation of our approach to the restricted domain SIS.

For evaluation, some queries are taken on student database of a college and analyze the generated answers but limitation is that our system unable to answer the question that belongs outside its domain.

QUERIES**:** 4 Events, each with 5 queries

Names:

- Write all names of students in engineering.
- Describe the name of students studying in CSE, ME and EE.
- List all the names of students in COE.
- Project names of studying people in a file.
- Names of student pursuing engineering.

SQL Query:

```
SELECT Student.names

FROM    Student

WHERE       BRANCH = "ME" ^
            BRANCH = "CSE" ^
            BRANCH = "EE" or
            COLLEGE = "COE"
```

2. Courses:

- Describe courses offered in the COE.
- Explain the studies areas of the students.
- Write all the courses covered by student.
- Illustrate fields in which students educated.
- Explain the subject taught by teachers in COE.

SQL Query:

```
SELECT Student.courses

FROM Student

WHERE       BRANCH = "ME" ^
            BRANCH= "CSE" ^
            BRANCH = "EE"
            and COLLEGE =
            "COE"
```

3. Grades:

- List the performance of all the students.
- Shows the result of all engineering students.
- Shows the result of all engineering students.
- List the grades of all students of CSE, ME and EE.
- Describe the grades of students of COE.

SQL Query:

```
SELECT Student.grades

FROM Student

WHERE       BRANCH = "ME" ^
            BRANCH= "CSE" ^
            BRANCH = "EE" or
            COLLEGE = "COE"
```

The primary reason for the wrong answer is the failure of invalid topic rejection. It is due to the insufficient of domain ontology data. An error was caused by the flaw of our keyword-based query frame decision approach. The system cannot answer to that topic and can't be returned, but our keyword based approach failed to make an appropriate query. To solve this problem, more sophisticated semantic representation, rather than the sequence of keywords, is required for the question.

5. RELATED WORKS

In this section, we compare our system with other SIS related approaches and briefly describe the distinctive characteristics of our system. Open-domain SIS systems in SI track mostly extracts answers from unstructured documents. In the contrast, our system extracts answers from table, which are selected by us, because our system aims to achieve high precision with the sacrifice of the coverage of questions. Natural language front ends for databases [4] and our system handle user questions similarly. However, our system has information extraction part that makes the database be updated regularly and automatically. Moreover, our system returns natural language responses to users. The system covers much wider domain of questions than ours, however, it seems that the system returns more wrong answers than ours, because the answers are extracted only from semi-structured documents. The Parser parses the question with the TINA language understanding system [5] and generates SQL and natural language answer with the GENESIS system [6]. The generated answer is synthesized with the SIS system [7].

6. CONCLUSION

This paper describes the SIS for restricted domains. To be practically used, our system tries to achieve high precision at the sacrifice of question coverage. A domain-specific ontology and query frames are prepared for the question analysis. By restricting the coverage of questions, our system could achieve relatively high precision. Much work is left for future work. First, the domain are expanded for the system. A domain classifier will be added to the SIS system. Domain dependent resources (query frames, ontology containing domain-dependent information, and etc.) and domain independent resources (linguistic resources , and ontology for domain-independent information) will be separated to allow easier domain expansion. Second, the size of ontology will increase to cover more question types. From the experimentation, it is realized that a larger ontology for sis is necessary to classify a question correctly. It seems more query

frames are necessary for more proper answers to the users' requests.

7. REFERENCES

[1] Ellen .M. Voorhees. 2004. Overview of the TREC 2003 question answering track. In Proceedings of the 12th Text Retrieval Conference

[2] M. Light, A. Ittycheriah, A. Latto, and N. Mac- Cracken. 2003. Reuse in question answering: A preliminary study. In New Directions in Question Answering: Papers from the 2003 AAAI Symposium, pages 78–86.

[3] Fellbaum. 1998. WordNet: an Electronic Lexical Database. The MIT Press.

[4] [4] Copestake and K. Spark Jones. 1990. Natural language interfaces to databases. The Knowledge Engineering Review, 5(4):225–249.

[5] B. Katz. 1997. Annotating the World Wide Web using natural language. In Proceedings of the 5th RIAO conference on Computer Assisted Information Searching on the Internet.

[6] S. Seneff. 1992. Tina: A natural language system for spoken language applications. Computational Linguistics, 18(1):pp. 61–86.

[7] L. Baptist and S. Seneff. 2000. Genesis-II: A versatile system for language generation in conversational system applications. In Proceedings of International Conference on Spoken Language Processing.

Appendix: Example of Student Information Systems

Q1: Write all names of students in engineering.

A1**:** The **Names** of students are as follows:

Names
John
Marry
Jessy
Rock
Jack
Bruce

Q2: Describe courses offered in the COE.

A2**:** The courses of students are as follows:

Courses
DB, DS
DB, DS
HMT, T
HMT, T
CS, CT
CS, CT

Q3: Explain the residential locations of the students.

A3: The system cannot answer to that topic.

Q4: List the performance of all the students.

A4: The Grades of students are as follows:

Grade
A+
A++
B
C
A
C

Q5: What is the ultraviolet index?

A5: The system cannot answer to that topic.