CPS 296.1 (Spring 2012): Project in Computational Journalism

Natural Language Interfaces to Databases
Roadmap

- Introduction, history, advantages and disadvantages
- Linguistic problems
- Architectures
- Portability
- Other issues
Introduction

> Who works on 3 projects?
B. Vandecapelle, C. Willems, D. Sedlock, J.L. Binot, L. Debille, ...
> Which of them are project leaders?
D. Sedlock, J.L. Binot
> Documents describing their projects?
BIM_LOQUI: "The Loqui Nlidb", "BIM_LOQUI"
MMI2: "Technical Annex"
> How many of these projects do not finish before 1994?
2
BIM_LOQUI, MMI2
> Are they led by JLB or DS?
The former.

• Aggregation, discourse context, predicting user intention, paraphrasing answers...
History

- NLIDB has been around since late 1960s (and longer than relational database systems)
- NLIDB research peaked in 1980s, but failed to gain commercial acceptance—why?
  - Simpler interfaces took off: graphical and form-based
  - More precise interfaces became easier: SQL
  - Intrinsic problems with NL interfaces
- How would DBMS development affect NLIDB?
  - Object-oriented DBMS? More complex structures?
  - Pervasiveness pushing the envelope of usability?
Advantages

• No artificial language (really?)
  • Limited subset of NL is still artificial!
• Better for some questions
  • E.g., “which company supplies every department?”
• Discourse
Disadvantages

- Linguistic coverage not obvious
  - Understood: “bordering Baltic and bordering Sweden”
  - Not understood: “bordering Baltic and Sweden”

- Linguistic vs. conceptual failures
  - Did the system not understand “salary” (vs. “wage”) or did it not have any salary information?

- Users assume intelligence
  - “Who will get a raise?”

- Inappropriate medium—NL is verbose and ambiguous

- Tedious configuration
Linguistic problem: Modifier attachment

- “[employees in the company] with a driving license”
- “employees in [the company with a driving license]”
  - Domain knowledge, or
  - Heuristics (e.g., “most right association principle”)
- “[employees in the division] making shoes”
- “employees in [the division making shoes]”
  - Ask user to clarify
  - Show both interpretations
Linguistic problem: Quantifier scoping

- “Has every student taken some course?”
  - $\forall$student $\exists$course (student has taken the course)
  - $\exists$course $\forall$student (student has taken the course)

- Heuristics (e.g., left-to-right)

- Assign numeric strength to each determiner (e.g., say $\text{strength(“every”) > strength(“some”)}$)
Linguistic problem: Conjunction and disjunction

- “List all applicants who live in California and Arizona”
  - We should detect that user obviously means “or” here
  - Is the test “query happens to return empty result”?
- “Which minority and female applicants know Fortran?”
  - Ask user to clarify
  - Show both interpretations
Linguistic problem:

Nominal compound

- What does “research” mean in “research department” vs. “research system”?
- What does “big” mean in “big department”?
  - Require all legal noun-noun and adjective-noun compound to be declared during a configuration step
Linguistic problem: Anaphora

> Who leads TPI?
E. Feron

> Who reports to him?
C. Leonard, C. Willems, E. Bidonnet, P. Cayphas, J.P. Van Loo

> What do they work on?

<table>
<thead>
<tr>
<th>project</th>
<th>worker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docdis</td>
<td>C. Willems</td>
</tr>
<tr>
<td></td>
<td>J.P. Van Loo</td>
</tr>
<tr>
<td></td>
<td>P. Cayphas</td>
</tr>
<tr>
<td>Euros</td>
<td>C. Leonard</td>
</tr>
<tr>
<td></td>
<td>C. Willems</td>
</tr>
<tr>
<td></td>
<td>E. Bidonnet</td>
</tr>
</tbody>
</table>

> Which of these are leaders?
J.P. Van Loo

• Context, domain knowledge
Linguistic problem:

Elliptical sentences

- What is the price of the three largest single port fixed media disks?
- What is the speed of the three largest single port fixed media disks?
- What is the speed of the two smallest single port fixed media disks?
- What is the price of the two smallest single port fixed media disks?
- What is the price of the two smallest fixed media disks with two ports?
- What is the speed of the two smallest fixed media disks with two ports?

VS.

- What is the price of the three largest single port fixed media disks?
- speed?
- two smallest?
- How about the price of the two smallest?
- Also the smallest with two ports?
- speed with two ports?

• Maintain a discourse model
Linguistic problem:
Extragrammatical utterances

- “whats raynolds departments name”
- “What is the crago of the Orient Clipper?”
  - Forgive errors
  - Automatically make/suggest corrections
  - Techniques have parallels in web searches
Architecture:

Pattern-matching systems

<table>
<thead>
<tr>
<th>COUNTRIES_TABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COUNTRY</td>
</tr>
<tr>
<td>France</td>
</tr>
<tr>
<td>Italy</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

pattern: ... “capital” ... <country>
action: Report CAPITAL of row where COUNTRY = <country>

pattern: ... “capital” ... “country”
action: Report CAPITAL and COUNTRY of each row

- May work well, but can lead to bad failures that are difficult to correct
Architecture:
Syntax-based systems

- Parse the input into a syntax tree
- Apply rules to map the syntax tree to a query
- Mapping rules are difficult to write and can be brittle
Architecture:

**Semantic grammar systems**

- Parse the input into a semantic parse tree
  - Nodes encode knowledge of the domain
- Apply rules to map the parse tree to a query

- Semantic grammars are costly to develop and maintain
- Difficult to port to other knowledge domains
Architecture:
Intermediate representations

- Careful division between NL and DB, and between domain-specific and generic components
  - Why?
  - Portability / extensibility!
Logical query

```
answer([Capital, Country]):-
    is_country(Country),
    borders(Country, greece),
    capital_of(Capital, Country).
```

- Captures the query at the logical level, independent of
  - The target database query language
  - The particulars of the database schema
- Supports multiple database systems with different query languages—even simultaneously!
Domain-dependent knowledge

- Lexicon: words that may appear in user’s questions, together with logic expressions
  - E.g., X is a “country” → is_country(X)
- World model: a hierarchy of classes of world objects, as well as constraints on the types of arguments to logic predicates
  - E.g., managers are employees; employees have salaries but clients don’t
  - Helps resolve some ambiguity
- Mapping to database information: how logic predicates relate to database objects
  - E.g., is_country → countries_table
Architecture:
Response generation

- Translate encoded information (e.g., id to names)
- Provide alternative Q&A when an answer is empty or not very informative
  - E.g., answering “yes” to “is there a flight to Athens?” is not very useful
- Translate queries back to NL to show to user to help avoid misunderstanding
Portability

- Across knowledge domains
  - Ask programmer, knowledge engineer, database administrator for input in advance, or
  - Ask user to teach the system on demand
- Across database systems
- Across natural languages
- Across hardware and software platforms
- Factoring and modularity help!
Other issues:

Restricted NL input

- Restrict syntax

  what are
  the names, ids, and categories of the employees
  who are assigned schedules
  that include appointments
  that are executions of orders
  whose addresses contain ‘maple’ and
  whose dates are later than 12/15/83 and
  whose statuses are other than ‘comp’

  what are
  the addresses of the appointments
  that are included in schedules
  whose call times are before 11:30 and
  that are executions of orders
  whose statuses are other than ‘comp’

- Use menus to limit vocabulary and structure
Other issues:

**NLIDB as intelligent agents**

- Support reasoning on top of raw data, e.g., “who are likely to develop lung disease?”
  - Requires pre-defined rules to compute derived data
  - Shouldn’t this be DBMS’s job instead of NLIDB’s?
- Reason about user’s goals
  - “Do American Airlines have a night flight to Dallas?”
  - “No, but they have one to Miami.”
  - “No, but they have a morning one.”
  - “No, but Delta has one.”
- Dialog systems: the logic of what the system should ask from the user can be coded as rules as well
Other issues:
NL specification of updates

- Can maintain a user view of the database to help resolve ambiguous updates

<table>
<thead>
<tr>
<th>employees_table</th>
<th>departments_table</th>
</tr>
</thead>
<tbody>
<tr>
<td>employee</td>
<td>salary</td>
</tr>
<tr>
<td>Adams</td>
<td>3000</td>
</tr>
<tr>
<td>White</td>
<td>3500</td>
</tr>
<tr>
<td>Brown</td>
<td>2500</td>
</tr>
<tr>
<td>Smith</td>
<td>2500</td>
</tr>
</tbody>
</table>

The following dialogue takes place:

- *List the employees and their managers.*

<table>
<thead>
<tr>
<th>employee</th>
<th>manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>Fisher</td>
</tr>
<tr>
<td>White</td>
<td>Baker</td>
</tr>
<tr>
<td>Brown</td>
<td>Jones</td>
</tr>
<tr>
<td>Smith</td>
<td>Jones</td>
</tr>
</tbody>
</table>

- *Change Brown’s manager from Jones to Baker.*

Done.

- *List the employees and their managers.*

<table>
<thead>
<tr>
<th>employee</th>
<th>manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams</td>
<td>Fisher</td>
</tr>
<tr>
<td>White</td>
<td>Baker</td>
</tr>
<tr>
<td>Brown</td>
<td>Baker</td>
</tr>
<tr>
<td>Smith</td>
<td>Baker</td>
</tr>
</tbody>
</table>

Seeing the unexpected change to view, the system would prefer updating Brown’s department to marketing.
Other issues

• Meta-knowledge questions
  • What data/query constructs are supported
  • Modal questions, e.g., “can people under 18 vote?”

• Temporal questions
  • Can the system under temporal constructs in NL?
  • Does the database manage historical information?

• Multimodal interfaces: merge NL with forms, GUI, ...

<table>
<thead>
<tr>
<th>put</th>
</tr>
</thead>
<tbody>
<tr>
<td>object:</td>
</tr>
<tr>
<td>destination:</td>
</tr>
</tbody>
</table>
Summary and thoughts

- Lots of interesting ideas (interfaces, interactions modes, heuristics, architectures) have been considered
- NLIDB is no longer as fashionable a research topic as it was in 1980’s
- As of the time of this article (1995), much of the work was still logic-based and rule-driven
  - Statistical machine learning has matured since then
- The need for better querying interfaces is still pressing
  - Some efforts on making querying easier for non-database programmers
  - But what about the general public?