quadri:
bumps in the road from language to data

presented by
richard waldinger

joint work with cleo condoravdi
danny bobrow, kyle richardson, and amar das
9 march 2012
why do we need logic?

Want to distinguish between

A patient does not have a regimen with AZT.

and

A patient has a regimen. The regimen does not have AZT.

Go
axiomatic subject domain theory

- defines concepts in queries.
- describes constructs in database.
- introduces the background knowledge that bridges the gap between them.
SNARK: theorem proving

- full first order logic: resolution
- equality reasoning: paramodulation, rewriting.
- ontology reasoning: sorted logic.
- temporal reasoning: allen temporal interval calculus, date and time arithmetic.
- answer extraction.
- procedural attachment...
  ➔ created by Mark Stickel at SRI
procedural attachment

- symbols in domain theory linked to procedures:
  - data base look-up
  - other computations
- when symbol appears in search, corresponding procedure is invoked.
- results of computation introduced into proof search.
  - virtual extension of theory
derived objects

- entity allowed in query.
- defined in domain theory.
- not represented explicitly in the database.
  - duration (finish-time - start-time)
  - “treatment change episode” (tce).
Show me patients on AZT.

there exists a patient14 such that
there exists a regimen15 such that
there exists a azt13 such that
patient14 is a patient and
patient14 has regimen15,
regimen15 has azt13 and
azt13 is azt
donkey anaphora

- *a patient has a regimen with azt.*
  
  exists(\textit{?patient}, \textit{?regimen})
  
  patient-has-regimen(\textit{?patient}, \textit{?regimen}) \&
  
  regimen-has-drug(\textit{?regimen}, azt)

- *the regimen is of at least 24 weeks.*
  
  duration(\textit{?regimen}) \geq \text{weeks}(24)
  
  - note “the regimen” is outside of the scope of the quantifier for \textit{?regimen}.
  
  - treated by squeezing the new condition inside the scope of the quantifier.
donkey anaphora

- *a patient has a regimen with azt.*
  
  \[
  \text{exists}(\text{?patient}, \text{?regimen}) \hspace{1cm}
  \text{patient-has-regimen}(\text{?patient}, \text{?regimen}) \& \\
  \text{regimen-has-drug}(\text{?regimen}, \text{ azt})
  \]

- *the regimen is of at least 24 weeks.*
  
  \[
  \text{duration}(\text{?regimen}) \geq \text{weeks}(24)
  \]
  
  - note “the regimen” is outside of the scope of the quantifier for ?regimen.
  
  - treated by squeezing the new condition inside the scope of the quantifier.
cardinality quantifiers

• **the regimen has a least 2 drugs.**
  
  \[ \exists \geq 2 \ ? \text{drug} \]
  
  \[ \text{regimen-has-drug}(\text{?regimen}, \ ?\text{drug}) \]

• **translated into**

  \[ \exists \ ?\text{drug1} \]
  
  \[ \text{regimen-has-drug}(\text{?regimen}, \ ?\text{drug1}) \& \]
  
  \[ \exists \geq 1 \ ?\text{drug} \]
  
  \[ \text{regimen-has-drug}(\text{?regimen}, \ ?\text{drug}) \& \]
  
  \[ ?\text{drug} \neq ?\text{drug1} \]

• **or**

  \[ \text{card(drugs-of regimen}(\text{?regimen}) \geq 2 \]
bridge anaphora

- find a patient with a tce.
  (failing regimen) • (salvage regimen)
- The patient has a high viral load 24 weeks before the baseline.
  ■ what is the "baseline"?
evaluation

- SweetInfo: provides graphical answers to queries....
- evaluation replicates a discovery from the literature.
- adding a box to the HIV database treatment change episode page.
What patients had a high viral load after 24 weeks on a regimen with RTV?
metaquadri

● replace hiv theory with arbitrary theory.
● introduce vocabulary.
● pass sort structure back into parser to remove ambiguities.
● allow new axioms to be introduced as declarative English sentences.
what’s the problem?

● provide access to novice users—physicians and researchers.
● a single query can require access to multiple databases.
● answers may need to be deduced or computed.
● database languages (e.g. sql) require specialized expertise.
how is this different from google, watson, siri, etc.?

- understanding of question.
- precise answers to questions.
- understanding of subject domain.
- focused subject domain.
our approach

- ask questions in English.
- translate into a logical form.
- reason in a theory of the subject domain (HIV treatment).
- allow the reasoner to access appropriate databases.
the quadri team

natural language—parc.
  cleo condoravdi (now stanford csli)
  dan bobrow
  kyle richardson (now university of stuttgart)

reasoning—sri
  richard waldinger
  tomer altman

database and hiv expertise—stanford
  amar das
  robert shafer
  soo-yon rhee

funding: NIH National Library of Medicine
hiv ontology

- patients
- regimens
- drugs
- viral loads
- mutations (genetic tests)
  - stanford hiv database
    - shafer, rhee
example

● What patients on azt exhibited a high viral load?
● parc’s xle translates into logical form (a theorem).
  \[\text{exists(?patient)[patient-has-regimen...}\]
● sri’s snark proves theorem and extracts answer from proof.
  \[\text{patient-id(605) ...}\]
● stanford’s hiv-db (and others) provides data.
axiomatic hiv theory

- defines concepts in query language.
- describes capabilities of data sources.
- provides background knowledge to link them together.
  - sorted axiomatic theory.
  - independent of any one data source.
- includes ontology.
sample axiom

high(viral-load, ?measurement) 
⇔ log(?measurement) ≥ 4

- i.e., a viral load measurement is high if and only if its log is greater than or equal to 4.
challenges in use of natural language

- language of query different from language of data source.
  - qualitative vs. quantitative
  - approximate vs. precise
- English is highly ambiguous.
- Query may be expressed as a sequence of questions.
mapping english to symbols

patients on azt ⇒
  patient-has-regimen(?patient, ?regimen) &
  regimen-has-drug(?regimen, azt)

● domain dependent.
● ?regimen implicit.
ambiguity

- **patients had a regimen with azt.**
  
  azt modifies regimen (correct) or azt modifies had (wrong).

- **I had a martini with an olive** vs. **I had a martini with Olivia.**

  (A martini can have an olive but cannot have Olivia.)
approaches to ambiguity

- use ontology to discard syntactically plausible but semantically meaningless readings.
- e.g., azt is a drug
  - a regimen can have azt.
  - azt cannot have a regimen
domain knowledge reduces ambiguity

Find patients who had a high viral load after 24 weeks on a regimen with azt.

- 62 readings without subject domain knowledge.
- 1 reading with subject domain knowledge.
logical form

Find patients who had a high viral load after more than 24 weeks on a regimen with azt.

ex(?pat, ?reg)
  patient-has-regimen(?pat, ?reg) &
  regimen-has-drug(?reg, azt) &
ex(?viral-test, ?time-point)
  patient-has-test(?pat, ?viral-test) &
  test-has-time(?viral-test, ?time-point) &
  test-has-result(?viral-test, ?test-result) &
  submeasurement(viral-load, ?test-result, high) &
ex(?time-interval)
  duration(?time-interval) ≥ 24*weeks &
  start-time(?time-interval) = start-time(?regimen) &
  finish-time(?time-interval) = ?time-point.
playback

- logical form(s) translated back into unambiguous (if clunky) English.
- user may select among alternatives.
- user may rephrase query if necessary.
playback example

- English: Find patients who have no regimens with azt.

- Playback:
  there exists a patient1 such that
  for all regimen2's,
  patient1 is a patient and it is not so that
  patient1 has regimen2 and
  regimen2 has azt
Theorem proving: SNARK

- Automatic first-order logic.
- Includes ontology reasoning.
- Answers to queries extracted from proof.
- Special procedures for temporal reasoning.
- Procedural attachment.
procedural attachment

- symbol in theory linked to
  - access of a table in data source.
  - other procedures
- when the symbol occurs in the proof search, the procedure is invoked.
- result of the procedure is introduced into the proof.
- axiomatic theory virtually extended.
  - e.g. patient-has-regimen(patient17, ?regimen)
procedural attachments to multiple data sources

- patient-has-regimen, patient-has-test
- the stanford hiv drug resistance database.
- other american and european sources planned.
display answers

- multiple proofs: multiple answers.
- user may request visual display of answers.
- SweetInfo project (Stanford) provides visual display of HIV data.
- evaluation of Quadri anticipated using SweetInfo test questions.
What patients had a high viral load after 24 weeks on a regimen with RTV?
explanation

- axioms and procedural attachments from the proof are used to construct an English paragraph that explains and justifies the answer and provides the provenance of the data invoked.
Sample Explanation

A patient has a high viral load if the log of the viral load is at least 4. The duration of a time interval is the difference between its finish point and its start point. ....

Patient 378 was on a regimen that started 29 August 1993. Patient 378 had a viral load of log 5 on 30 November 1999. ...

English transcriptions of

- axioms
- results of procedural attachments.
complex queries: multiple questions

- Find patients who exhibited m184v.
- The patients were on azt.
- The patients had a high viral load after more than 24 weeks on the regimen.
testing

• allow access to quadri via the stanford hiv database-the quadri box!
future work

- expressing new knowledge in English.
- adaptation to new subject domains
  - breast cancer
- providing health care information to patients.
new knowledge: high viral load

- English: A viral load is high if and only if the log of the viral load is greater than or equal to 4.
- Logic:
  \[ \text{high(viral-load, ?measurement)} \iff \log(?\text{measurement}) \geq 4 \]