Issues in Question Answering

William A. Woods

Sun Microsystems Laboratories

June 9, 2006

Announced Workshop Issues

- Robustness – used by real users outside of a lab
- Portability – generalizability to new domains
- Interaction – not just isolated questions
- Cognitive Augmentation – relieve the user of some burdens
- User Needs – different kinds of users with different needs
- Evaluation – classical quantitative metrics of precision and recall are not sufficient

More Workshop Issues

- Models of Domain
- Models of Dialog
- Data Presentation

I would add:

- Importance of Collateral Knowledge
- Issues of Paraphrase Variability and Generality
- Models of Human Problem Solving States
- Models of Knowledge of Others

Some Historical Benchmarks

- 1963
  - Chomsky et al – Baseball
  - Simmons et al – Protosynthex
- 1964
  - Black – A Deductive Question Answering System
  - Bobrow – STUDENT
- 1968
  - Woods - Semantics for a Question Answering System
  - Winograd - SHRDLU

More Historical Benchmarks

- Woods et al – LUNAR – 1970+
- Carbonnel & Collins – SCHOLAR -- 1973
- Perrault, Cohen, Allen – Formalized Speech Acts
- Woods & Cohen et al – TRIPSYS – 1976
- Woods & Makhoul et al – HWIM – 1976
- Reichman – conversational moves – 1981
- Reiter – generating descriptions – 1990
- Planpower – Financial Planning Reasoning

More Recent Experiences

- Woods – Precision Content Retrieval – 1991+
  - Robustness
  - Portability
  - Interaction
  - Cognitive Augmentation
  - User Needs
  - Evaluation
  - Collateral Knowledge
  - Data Presentation

The Baseball System (1963)

Question: What teams beat the Tigers 5-4 in May?

Spec List:

(a) SUBLIST = { ([WINNING] TEAM = ?, SCORE = 5)
([LOSING] TEAM = TIGERS, SCORE = 4) }

(b) MONTH = MAY

Question: How many times did the White Sox score 5 runs against the A's after July 4?

Spec List:

(a) # ? TIMES = –

(b) SUBLIST = { (TEAM = WHITE SOX, SCORE = 5)
(TEAM = ATHLETICS) }

(c) DATE = [AFTER] JULY 4

Question: How many times did the White Sox yield 5 runs to the A's after July 4?

Spec List:

(a) # ? TIMES = –

(b) SUBLIST = { (TEAM = WHITE SOX)
(TEAM = ATHLETICS, SCORE = 5) }

(c) DATE = [AFTER] JULY 4

Protosynthex – 1963

- Answers questions from encyclopedia articles
- Similar to current open-ended QA Paradigm
  - Indexes articles by their “root-form” content words
  - Uses synonym dictionary and “complex intersection logic” to score sentences and paragraphs that most resemble the question
  - Parses the retrieved sentences and the question, using a dependency grammar and human interaction
  - Matches parse of query to parses of sentences
  - Answer is any phrase that matches the query word

Protosynthex – example query

What do worms eat?

output:

worms = worms

eat = eat

what = grass

what = their way

what = through the ground

(The system was actually more sophisticated than this makes it sound – e.g., semantic filtering of answers)

Fisher Black – 1964

A Deductive Question Answering System

- Queries and data expressed in first-order logic
- No natural language involved
- Apparently the first to take quantifiers seriously
- Other people worked on English-to-Logic translation
  - e.g., Williams (1956), Bohnert (1963), Darlington (1964)

Bobrow's Student (1964)

```
(THE PROBLEM TO BE SOLVED IS)
(THE DISTANCE FROM NEW YORK TO LOS ANGELES IS 3000 MILES .
IF THE AVERAGE SPEED OF A JET PLANE IS 600 MILES PER HOUR ,
FIND THE TIME IT TAKES TO TRAVEL FROM NEW YORK TO LOS ANGELES
BY JET .)


(THE EQUATIONS TO BE SOLVED ARE)

(EQUAL G02517 (TIME (IT / PRO) TAKES TO TRAVEL FROM NEW YORK
TO LOS ANGELES BY JET))

(EQUAL (AVERAGE SPEED OF JET PLANE) (QUOTIENT (TIMES 600 (MILES))
(TIMES 1 (HOURS))))

(EQUAL (DISTANCE FROM NEW YORK TO LOS ANGELES) (TIMES 3000
(MILES)))
```

Bobrow's Student (1964)

```
THE EQUATIONS WERE INSUFFICIENT TO FIND A SOLUTION

(USING THE FOLLOWING KNOWN RELATIONSHIPS)
((EQUAL (DISTANCE) (TIMES (SPEED) (TIME))) (EQUAL (DISTANCE)
(TIMES (GAS CONSUMPTION) (NUMBER OF GALLONS OF GAS USED))))

(ASSUMING THAT)
((SPEED) IS EQUAL TO (AVERAGE SPEED OF JET PLANE))

(ASSUMING THAT)
((TIME) IS EQUAL TO (TIME (IT / PRO) TAKES TO TRAVEL FROM NEW
YORK TO LOS ANGELES BY JET))

(ASSUMING THAT)
((DISTANCE) IS EQUAL TO (DISTANCE FROM NEW YORK TO LOS ANGELES))


(THE TIME IT TAKES TO TRAVEL FROM NEW YORK TO LOS ANGELES BY
JET IS  5 HOURS)
```

SHRDLU (Winograd 1968) explored

- Black box computer programs for the entire Question Answering process (just program it)
- Programmed implementation of Halliday's theory of Systemic Grammar
- Questions and actions in a simulated world
- Using a world model to resolve ambiguities
- Unwinding the goal stack for answering “why” questions

"Semantics for a Question Answering System" Woods 1968

- Introduced generalized quantifiers
- Extended logic with commands and actions
- Procedural semantics as theory of meaning
- Formal Meaning Representation Language
- Transformed English to MRL expressions
- Can execute or reason with MRL expressions
- Anaphoric reference to quantified noun phrases

Formal Meaning Representation Language

an extension of the Predicate Calculus with generalized quantifiers and imperative operators

(FOR <quant> <vbl> / <class> : <condition> ; <command>)

(FOR <quant> <vbl> / <class> : <condition> ; <condition>)

(TEST <condition>) (PRINTOUT <designator>)

Generalized Quantifiers with Type Functions and Restrictions

allow uniform treatment of different quantifiers in noun phrases

"Some tall men play basketball"

(FOR SOME X / MAN : (TALL X) ; (PLAY X BASKETBALL))

"All long flights to Boston serve meals"

(FOR EVERY X / (FLIGHT-TO BOSTON) : (LONG X) ;
(SERVE-MEAL X))

Compare with Mapping English to Classical Logic

needs different treatments for different quantifiers in noun phrases

"Some tall men play basketball"

(THERE EXISTS X) [(MAN X) AND (TALL X) AND
(PLAY X BASKETBALL)]

"All tall men play basketball"

(FOR ALL X) [((MAN X) AND (TALL X)) IMPLIES
(PLAY X BASKETBALL)]

Procedural Semantics

- Meaning is embodied in abstract procedures for determining referents, verifying facts, computing values (including truth values), and carrying out actions
- These procedures are built on computational operators -- cons, cond, loop, read, print, etc., and can include sensing and acting on the world
- This provides a principled connection between "mental" symbols and what they denote or mean (in a real world as well as in a modelled world)

Procedural Semantics

- permits a computer to understand, in a single, uniform way, the meanings of conditions to be tested, questions to be answered, and actions to be carried out
- permits a general-purpose system for language understanding to be used with different data bases that may have different representational conventions and different data structures

Reasoning with Meanings

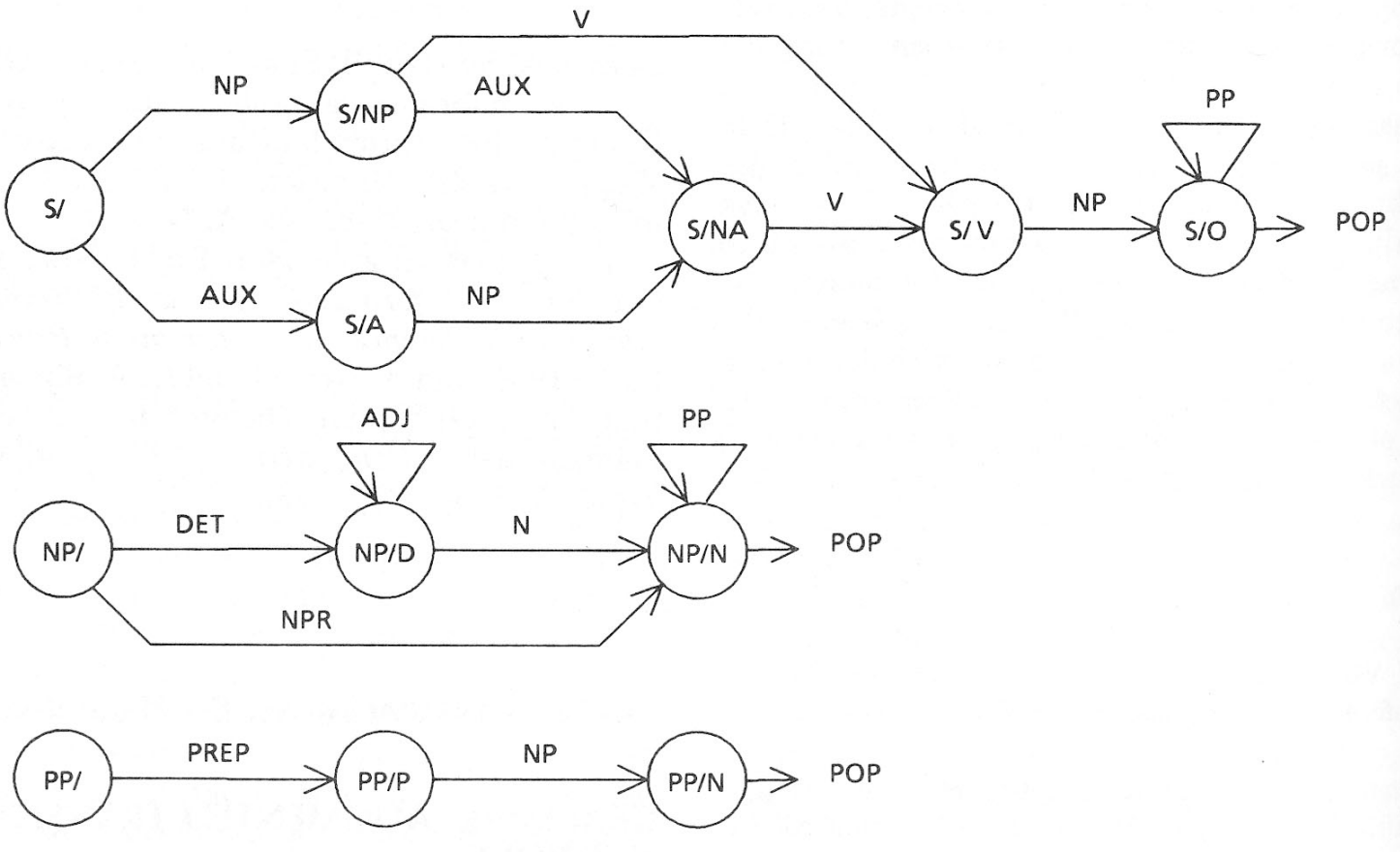
Procedural meanings are symbolic expressions

- simplest model -- simply execute the semantic interpretation as a program
- more general model -- you can do logical reasoning with it (e.g., optimize it first)

(FOR EVERY X / FLIGHT : (CONNECT X BOSTON CHICAGO) ; (PRINTOUT X))

(FOR EVERY X / (FLIGHT-TO CHICAGO) : (CONNECT X BOSTON CHICAGO) ; (PRINTOUT X))

ATN Grammar State Diagram



Answering Questions about Paths in an ATN:

- Involves quantifying over objects that don't preexist in the database
- Enumerating all possible paths (without restrictions like where they start or end or loop) and then restricting them would be inefficient
- Smart quantifiers can reason about the restrictions and then use a specialized enumeration function that's more efficient

Procedural Semantics in LUNAR (1971)

What is the average concentration of Aluminum in each breccia?

```
(FOR EVERY X5 / (SEQ TYPECS) : T ;
    (PRINTOUT (AVGCOMP X5
        (QUOTE OVERALL) (QUOTE AL2O3))))
```

Anaphora in LUNAR (1971)

Which samples contain ulvo-spinel?

S10044

S10045

S10060

S10084

Give me all Chromite analyses for those samples

I have 10 hits, do you want to see them?

Yes

...

Limitations in LUNAR (1971)

What is a breccia?

S10018

What is S10018?

S10018

LUNAR simply finds referents of referring expressions and gives their names

There is no model of the purpose behind the user's question or of different kinds of answers for different purposes

Describing Objects (Reiter, 1990)

What is Unix?

- an operating system
- an operating system with a powerful command line interface
- an operating system that manages the resources of the underlying computer hardware
- an operating system developed at Bell Labs in the early 1970's
- an operating system that provides a genuine multi-tasking, multi-user environment

Conversational Moves (Reichman, 1981)

Reichman studied real conversations between people discussing a variety of topics, including political discussions and technical explanations

She found consistent, regular structure across these diverse kinds of discourse

Conversational moves correspond to communicative goals

e.g., support, interrupt, challenge, illustrate, ...

She used an ATN grammar to characterize how a move depends on and changes the state of the discourse, using the ATN's ability to model deep relationships

Some Questions that Require Explanations – Why and How

- One kind: Subgoal of a higher goal (SHRDLU)
  - Why did you move the triangle?
  - To clear the top of the block
  - Why did you put the block in the box?
  - Because you told me to
  - (Basis of most Expert Systems explanations)
- Another: How will it benefit me? (PlanPower)
  - It will save you money

Causal Explanations require Models of others' Knowledge

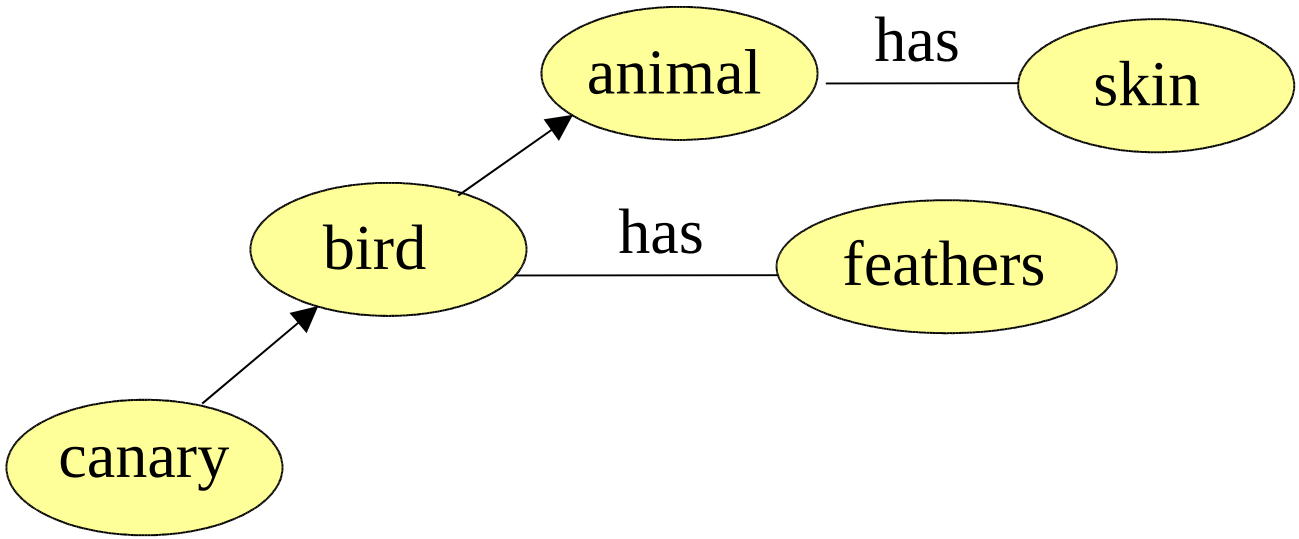
- Simplest case: Consider P implies Q
  - assume P is true and therefore Q is true
- The questioner asks “Why Q?” You answer:
  - Because P (if questioner knows that P imples Q)
  - Because P implies Q (if questioner knows that P)
  - Because P implies Q and P is true (if neither is known)
- In general, Q is at the end of an large tree of causes and assumptions – which is least known?

Modeling the Knowledge of Others: E.g., SCHOLAR — 1973

- SCHOLAR was a mixed initiative tutoring system to teach South American Geography
- It used a semantic network to store the knowledge to be taught
- The student could ask questions, and the tutor could ask the student questions
- The tutor kept track of what the student knew by annotating the semantic network

Associative Distance in a Semantic Network (Some things are known more directly than others)

"A canary has feathers" is recognized faster than
"a canary has skin" (because of an additional link in the inheritance chain?)



Collins's Lack-of-Knowledge Principle – better than CWA (some things are more certain)

- Question: “Does Bolivia export steel?”
- I know a lot about Bolivia's exports and if steel were a significant export, I would know it.
- I don't know about Bolivia exporting steel, so the answer is “No.”
- If I didn't know that much about Bolivia's exports, the answer would be “I don't know.”

Private, Public, Mutual, and Common Knowledge

Perrault, Cohen, and Allen (and others) have formalized a theory of communication based on beliefs, desires, intentions, speech acts, and mutual belief

Attunement: What knowledge do two communicators need to share in order to communicate efficiently?

How does a speaker diagnose and/or estimate his hearer's state of knowledge (and state of mind)?

How do the rules of conversation and propriety and the context restrict what you'd like to say and determine the way that what you say will be interpreted?

Representing Knowledge – Some Fundamental Issues

How does a reasoning system find relevant pieces of information and relevant rules of inference when it knows millions of things?

How does it acquire and organize millions of items of information?

How does it integrate new information with previously existing information?

How does it use its knowledge to impose structure on situations and decide what to do?

Requirements for Semantic Representation

We need a representational system to satisfy two requirements:

- expressively adequate to represent all of the necessary elements of natural language questions, commands, assertions, conditions, and designators
- structured to support semantic interpretation, retrieval, and inference

Links and Logic

Can we combine best of two traditions:

Logical Deduction

- rigorous, formal, but often counterintuitive
- expensive algorithms that match expressions, substitute values for variables, and invoke rules

Associative Networks

- associative, intuitive, but typically informal
- efficient algorithms that follow paths through links

Understanding Subsumption and Taxonomy (Woods, 1991)

Revisited goals of “What's in a Link” and KL-ONE

- Principled methodology for organizing knowledge for efficient use
- Intensional subsumption criterion is more expressive and more efficient than KL-ONE
- Quantificational tags capture semantics of links
- MSS and MGS algorithms find where new concepts belong in the taxonomy

MSS can find locations
for concepts that aren't there

```
cleaning
    automobile cleaning
!       automobile steam cleaning (1)
        automobile upholstery cleaning (1)
        automobile washing (1)
            car washing (1)
    industrial cleaning
        industrial steam cleaning (1)
    steam cleaning
!       automobile steam cleaning (2)
        industrial steam cleaning (2)
    upholstery cleaning
        automobile upholstery cleaning (2)
    washing
```

A Practical Application: NOVA Precision Content Retrieval

A tool for efficient answer finding

- Supports associative information access
- Finds specific passages of text in response to specific requests
- Helps people find specific information quickly

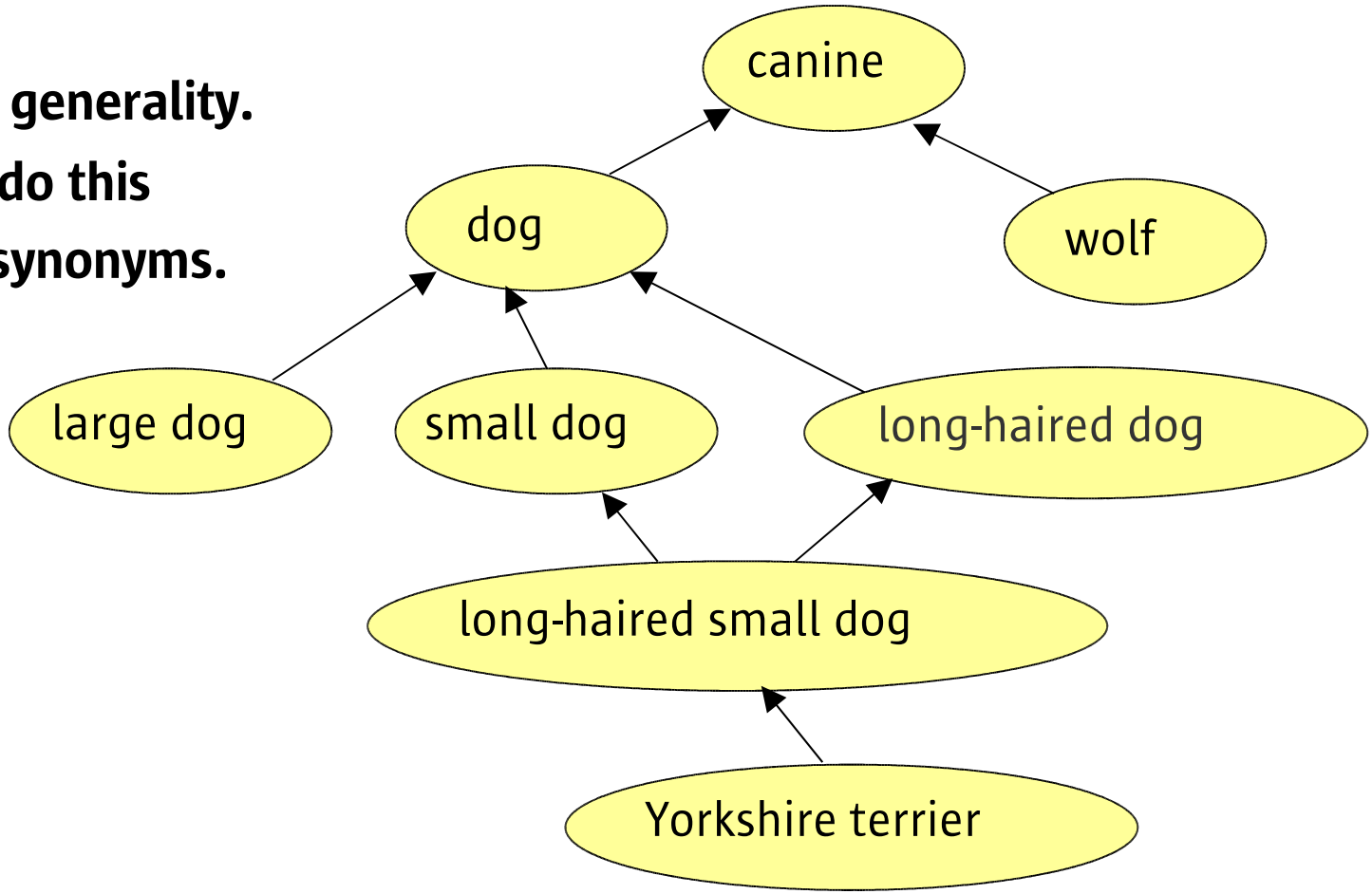
(and a lab to gain experience with large-scale natural populations of concepts and a good tool to support linguistic analysis of corpora as well)

Big Idea : taxonomic search

Search by generality.

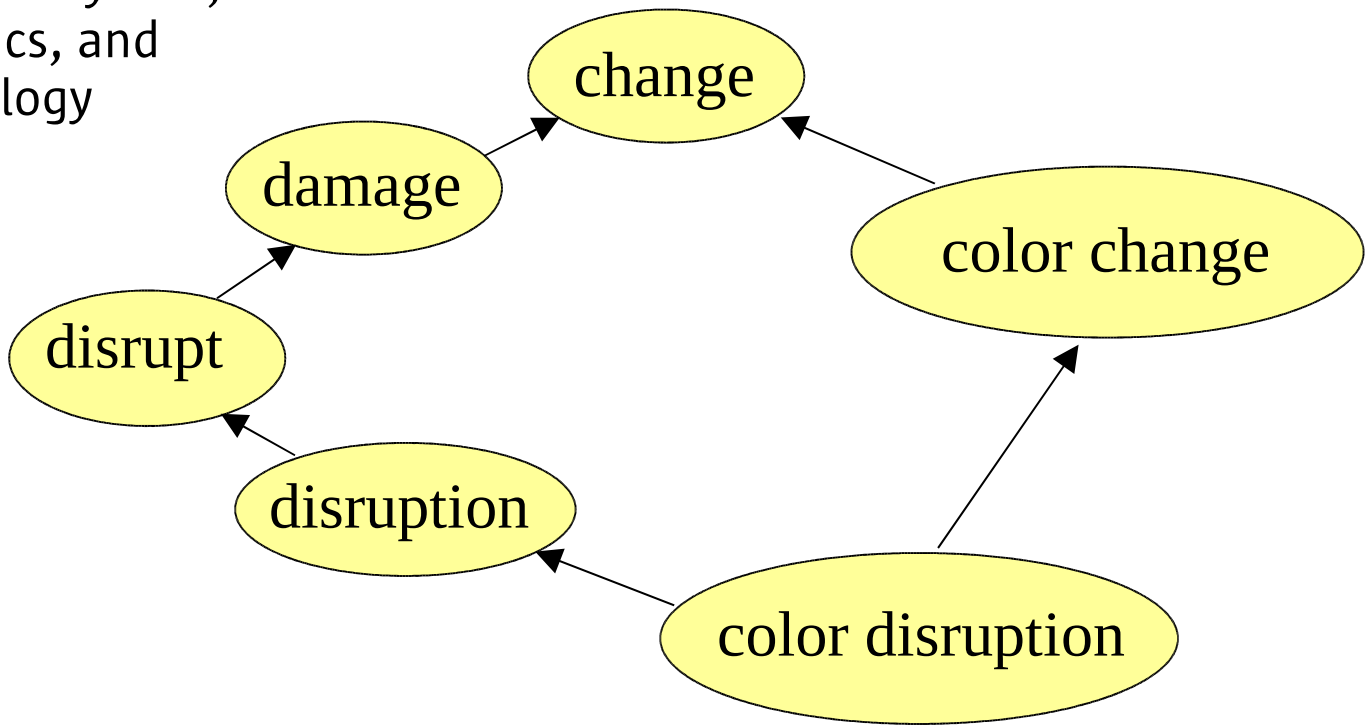
You can't do this

with just synonyms.



Structured Conceptual Taxonomy

integrates syntax, semantics, and morphology



Browsing in a Conceptual Taxonomy

BROWN FUR
  |-k- GRAY BROWN FUR
  |-k- RICH BROWN FUR
  |-k- WHITE-SPOTTED BROWN FUR

More general concepts:
  kind of BROWN COAT
    kind of COAT
      kind of SOMETHING
  kind of FUR
    kind of ANIMAL_COAT

```
BROWN COAT
        |-k- BRIGHT REDDISH BROWN COAT
        |-k- BROWN BLACK COAT
        |-k- BROWN COATS
        |       |-k- FAWN COATS
        |       |       |-v- (FAWN) COATS
        |       |
        |       |-k- REDDISH BROWN UPPER COATS
        |
        |-k- BROWN FUR ....
        |
        |-k- BROWN HAIR
        |       |-k- BROWN HAIRS
        |       |       |-k- REDDISH BROWN GUARD HAIRS
        |       |
        |       |-k- BROWN WOOL
        |       |       |-k- REDDISH BROWN WOOL
        |       |
        |       |-k- BROWNISH HAIR
        |       |-k- REDDISH BROWN HAIR
        |                 |-k- REDDISH BROWN GUARD HAIRS
        |                 |-k- REDDISH BROWN WOOL
        |
        |-k- BROWN-GRAY COAT
```

Passage Query Operator

Input query: black and white dog

Parsed query

```
QUERY: DOCUMENTS 1 (<passage> (<morph> "black") (<morph> "and") (<morph>
"white") (<morph> "dog"))
```

Search took: 2.895s

Hits 1 through 20 of 45525 shown

1. 100.0 Vandals Hit Pet Cemetery; Checkers' Plot OK

0.0 **black and white dog**

...he illegally accepted gifts from wealthy supporters. He cited the **black–and–white dog**, a supporter's gift to his family, as one contribution ... Jump to Hit

2.	99.9	World Not Just **Black and White** To **Dogs**, Study Concludes		
		0.8	**black and white dogs**	
			...brightness cues," he said. Besides disproving the popular notion that **dogs** see in **black and white**, the study gives insight into the eye's mechanisms for ...	Jump to Hit

Semantic Term Expansion

Input query: black and white dog

Parsed query

```
QUERY: DOCUMENTS 1 (<passage> (<expand> "black") (<expand> "and")
(<expand> "white") (<expand> "dog"))
```

Search took: 4.119s

Hits 1 through 20 of 47410 shown

1. 100.0 Vandals Hit Pet Cemetery; Checkers' Plot OK

   0.0 **black and white dog**

   ...he illegally accepted gifts from wealthy supporters. He cited the **black–and–white dog**, a supporter's gift to his family, as one contribution ... Jump to Hit

2. 100.0 Town Holds Daylong Seminar on Alzheimer's

   0.0 **black and white mongrel**

   ...love for an Alzheimer's sufferer, Connolly said. Jeff, a **black and white mongrel** from a Rhode Island animal shelter, has become the national ... Jump to Hit

Semantic Taxonomy

- sheepdog (10)
  - sheepdogs (3)
  - collie (27)
- showdog (3)
- swampdog (1)
- underdog (342)
  - underdogs (40)
- watchdog (659)
  - watchdoggery (1)
  - watchdogs (81)
- !n/dog (0)
  - !nc/dog (0)
    - !nc/mongrel (0)
    - airedale (4)
    - alaskan_dog (0)
    - beagle (35)
    - bulldog (38)
    - dachshund (14)
    - doberman_pinscher (0)

Weighted AND Operator

Input query: black and white dog

Parsed query

```
QUERY: DOCUMENTS 1 (<and> (and (<morph> "black") (<morph> "white"))
(<morph> "dog"))
```

Search took: 0.659s

Hits 1 through 20 of 294 shown

1. 37.4 LaserPhoto CPN1

0.0 **dogs blacks whites**

...Police used whips and **dogs** to disperse hundreds of **blacks**, including Archbishop Desmond Tutu, during mass protests Saturday at two **whites**–only beaches. Tutu was carried shoulder–high onto the first ... Jump to Hit

2 passages not shown

2.	37.0	After Years Of Protest, South African Beaches Open To All Races		
		0.0 **dogs blacks white**		
			...apartheid practices. As recently as August, police	Jump to

Precision Content Retrieval

- Robustness – a real deployed system at Sun
- Portability – has been applied to many domains
- Interaction – choose operators, follow links
- Cognitive Augmentation – semantic taxonomy
- User Needs – saves time, natural querying
- Evaluation – success rate, elapsed time, grades
- Collateral Knowledge – morphology, semantics
- Data Presentation – scores, words, passages

Good Synergy Between Human and Machine

- Computer finds and ranks passages
- Provides concise information to enable human to quickly spot answers
- Relies on human judgement to recognize answers
  - Doesn't know when it has an answer
  - Doesn't know what the answer is
  - Can't evaluate alternative answers
  - Can't combine information from different sources or judge partial answers

Beyond Passage Retrieval

Real Question Answering Requires
- understanding answer passages
  - (natural language parsing and semantic interpretation)
- and reasoning about them
  - (with background knowledge and scalable reasoning algorithms in a general reasoning architecture)

Evaluation Issues

- LUNAR was the first Question Answering System to publish performance data
  - 78% of questions successfully answered to live users at Second Annual Lunar Science Conference
  - 90% would have been answerable with minor fixes
- And first to qualify their significance
  - Not a controlled experiment
  - No single user asked more than one question
  - Predicted different character to follow-up questions

Evaluation Issues

- In TRIPSYS/HWIM (a travel budget management and trip planning system with spoken input), some interpretations of an utterance would make more sense than others, and that could depend on context (defined by a pragmatic grammar)
- Setup for classical evaluation was expensive
  - Each test utterance would require a complete contextual setup – including interpretations of all of the questions and commands leading up to it
- We never got far enough to do it

Evaluation Issues

- In evaluating precision content retrieval, we tried to address the fact that some passages were better than others by grading them:
  - A – completely answered the question
  - B – partially answered the question
  - C – didn't answer question but was related to it
  - D – didn't provide anything useful
- Assigning grades to passages is not a lot more difficult than judging relevance – in some ways it makes the job easier – it's not all or nothing

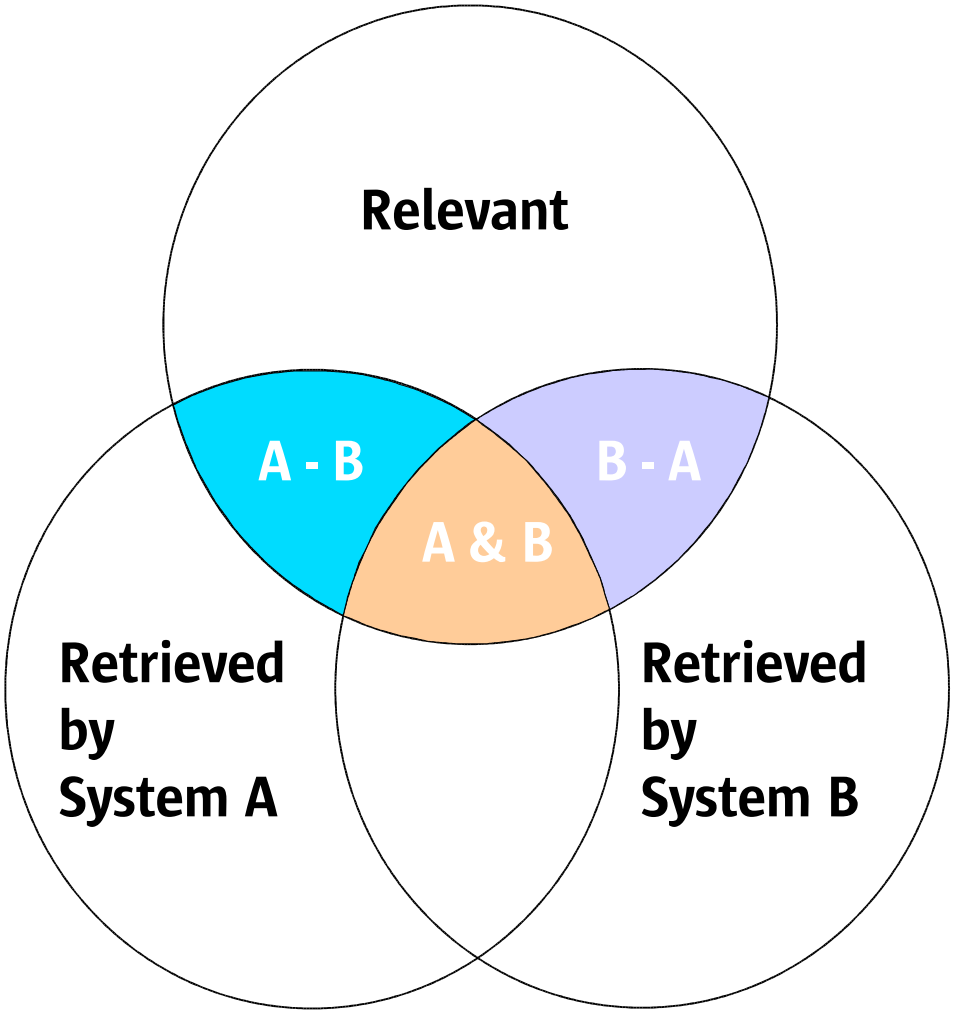
Evaluation Issues

- Classical Recall and Precision require ground truth data that we can't get:
  - For large collections, it's infeasable to determine the true relevant ground truth (if it ever was)
  - Current approximate methods are misleading
  - For dynamically constructed passage answers, there is not even a predefined set of things to consider
- Moreover, two systems with same recall don't necessarily retrieve the same things (or even the same kinds of things)

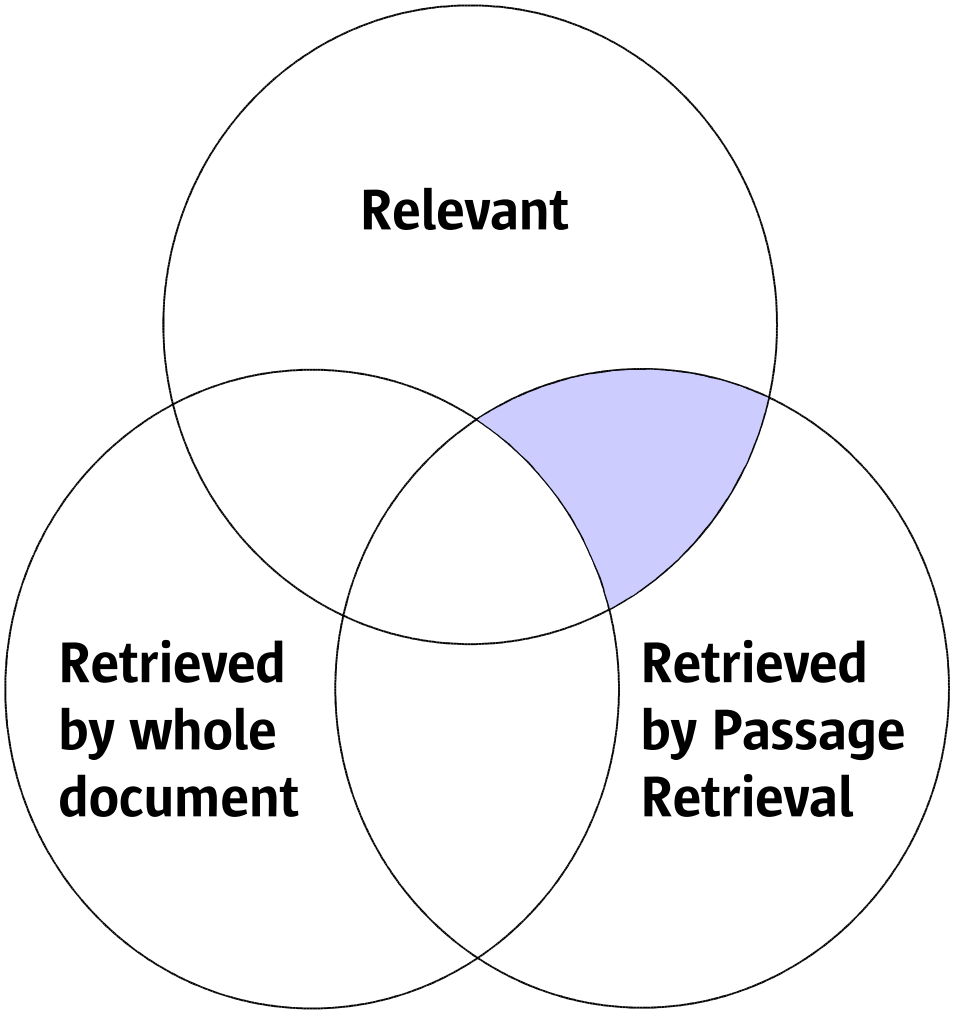
Evaluation Experience

- Specific-passage-based document retrieval complements summed weighted word counts:
  - For some tests, both methods had approximately the same success rates (at the document level)
  - But they didn't find the same documents
    - Passage-based is best for
      - Specific information
      - Short queries with relationships between words
    - Weighted word counts is best for
      - Topic-oriented queries
      - Relationships between words not important

Differential Retrieval



Passage Retrieval Benefits



Evaluation Conclusions

- Comparison of systems is not one dimensional:
  - For some questions additional answers after the first one don't matter, so classical recall is wrong measure
  - Success rate is what the user wants, but that conflates two factors:
    - Is the information in the database
    - Did the system find it if it was
  - Precision at 5 or 10 retrieved items is what the user wants, but that can be low because there aren't enough hits to find
  - We need to report and use a battery of measures

Evaluation Possibilities

- Rigorous psycholinguistic experiments are possible to measure cognitive load:
  - Kaplan and Stone investigated pschological reality of linguistic grammar formalisms
  - They measured the cognitive load at different points in the processing of a sentence:
    - Predicted more load where the grammar allowed more local possibilities
    - Measured reaction time for concurrent monitoring tasks
    - Found delayed reaction to stimuli at points where grammar predicted more load

Conclusion:

The range of open problems cannot be solved by mere application of one, or even a few, existing, well-understood techniques such as first-order logic, Bayesian statistics, and machine-learning algorithms, although these are clearly parts of the solution.

We need to forge new tools sufficient to the tasks.

E.g., models of mental states and knowledge of others