CIS-700

Spring 2020

Reasoning for Natural Language Understanding

Dan Roth

Computer and Information Science

University of Pennsylvania

Introduction Part III: Knowledge Representation and Reasoning – Classical View

This class

- Understand early and current work on Reasoning
  - (Learn to) read critically, present, and discuss papers
- Understand some of the difficulties in NLU from the perspective of reasoning
  - Conceptual and technical
- Try some new ideas

- How:
  - Presenting/discussing papers
    - Probably: 2 presentations each; 4 discussants
  - Writing a few critical reviews
  - “Small” individual project (reproducing);
  - Large project (pairs)
  - Tentative details are on the web site.

- Today: discuss first project
  - Content + Timetable
- Today: release list of papers
  - Timetable

- Machine Learning
  - 519/419
  - 520
  - Other?
- NLP
  - Yoav Goldberg’s book
  - Jurafsky and Martin
  - Jacob Eisenstein

- Attendance is mandatory
- Participation is mandatory

- **Time of class?**
- **Expectations?**

Reasoning

- The classical view of reasoning:
  - Deriving conclusions from a corpus of explicitly stored information, as a mean to solve a range of problems.
- An ideal reasoning system will produce:
  - All-and-only the correct answer to every possible query
  - Produce answers that are as specific as possible
  - Be expressive enough to permit any possible fact to be stored and and query to be asked
  - Be efficient
- Probably impossible for many reasons (?)
- Most of the classical research focused on tradeoffs:
  - As correct systems become more expressive, they can become less efficient
- This was studied both in the context of logic- and of probability-based reasoning.
- Less effort was devoted to connecting things to applications where reasoning is needed
  - Representation (and Mapping) – are these realistic?
  - Formulation – is it satisfactory?

Towards a Formulation

Deduction: Conclusion from given axioms (facts or observations)

All humans are mortal. *Socrates is a human.*	(axiom) (fact/ premise)
Therefore, it follows that Socrates is mortal.	(conclusion)

Abduction: Simple and mostly likely explanation, given observations

All humans are mortal *Socrates is mortal*	(theory) (observation)
Therefore, Socrates must have been a human	(diagnosis)

- ☐ Of these, abduction might be the most useful (?) in many situations.
- ☐ But, we need to formalize these.
- ☐ And, maybe think about the relations to Induction
- ☐ And, always ask, are these forms of reasoning sufficient?

Representation and Reasoning

- Propositions (p, q, …). Connectives ($\wedge$, $\neg$, …).
  - Implications: $\varphi \Rightarrow x$. Equivalences: $\varphi \Leftrightarrow x$.
- Reasoning ***semantics*** through entailment $\vDash$.
- Proof procedures $\vdash$ to ***compute*** entailment.

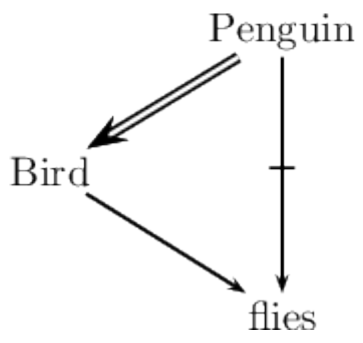
- Given formulas in KB and an input O, ***deduce*** whether a result R is entailed ($KB \cup O \vDash R$).
- Given formulas in KB and an input O, ***abduce*** an explanation E that entails O ($KB \cup E \vDash O$).

  - The question of how to compute deduction (and abduction) is also an interesting question here.

Credit: AAAI 2018 Tutorial: Integrating Learning into Reasoning, Juba and Michael

Non-Monotonic Reasoning

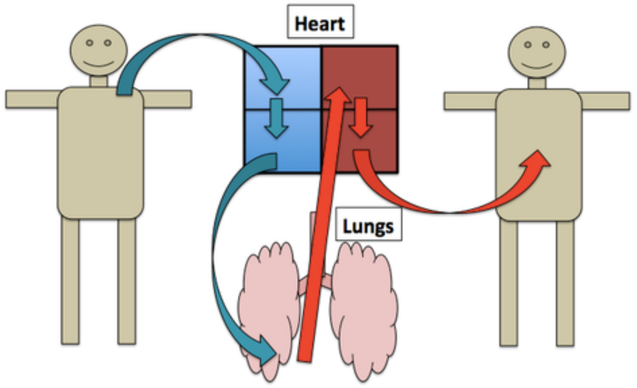
- Non-monotonicity typically viewed as property of ***extending input O for fixed KB***, and having result R become "smaller".
  - Birds fly
  - Tweedy is a bird; does Tweedy fly?
  - Tweedy is a penguin



- ☐ This is a problem to most formalisms
- ☐ Involving learning in the process provides ways to address these difficulties.

Analogy

- The heart is a pump
- Is this an important reasoning setting?



Quantitative Reasoning

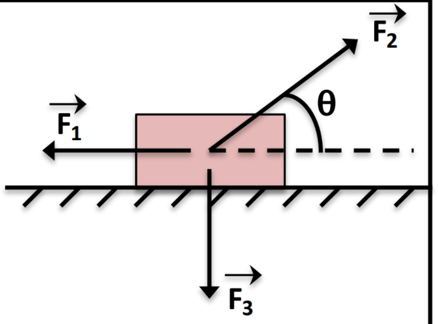
Example: The sum of two numbers is 111. One of the numbers is consecutive to the other number. Find the two numbers.

Example: Bill s father s uncle is twice as old as bills father. 2 years from now bill s father will be 3 times as old as bill. The sum of their ages is 92. Find Bill s age.

Example: The distance between New York to Los Angeles is 3000 miles. If the average speed of a jet place is 600 miles per hour find the time it takes to travel from New York to Los Angeles by jet.

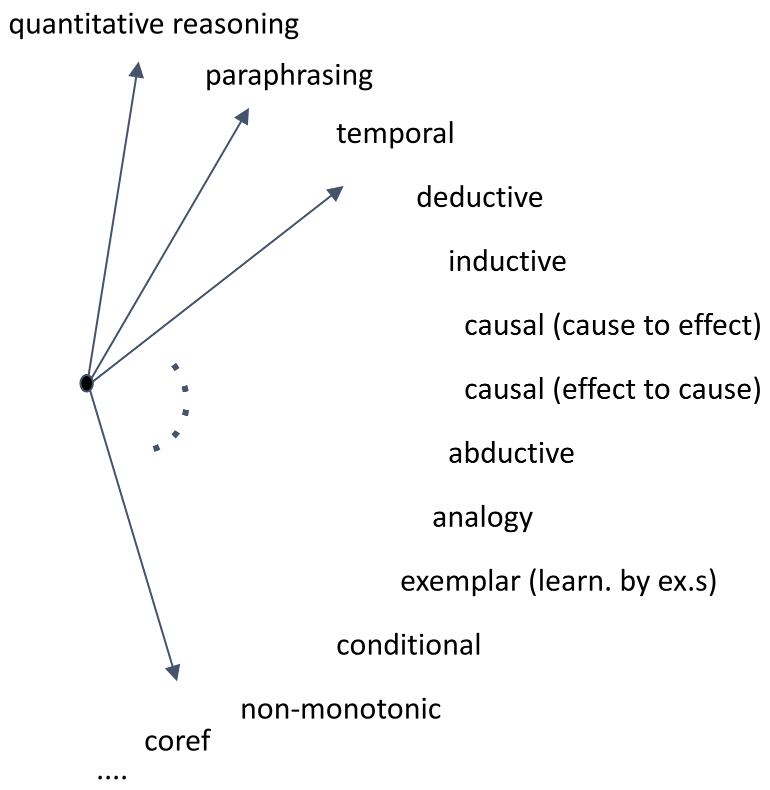
Example: Ram Emanuels' campaign contributions total that of all his competitors together.

The figure shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are $F1$ = 5.00N, $F2$ = 9.00N, and $F3$ = 3.00N, and the indicated angle is θ = 60.0° . During the displacement, what is the net work done on the trunk by the three forces?



The many faces of reasoning

- Reasoning is often studied in a very narrow sense.
  - But probably has many forms
  - Realistic examples typically span multiple reasoning aspects.



The many faces of reasoning

Abductive reasoning

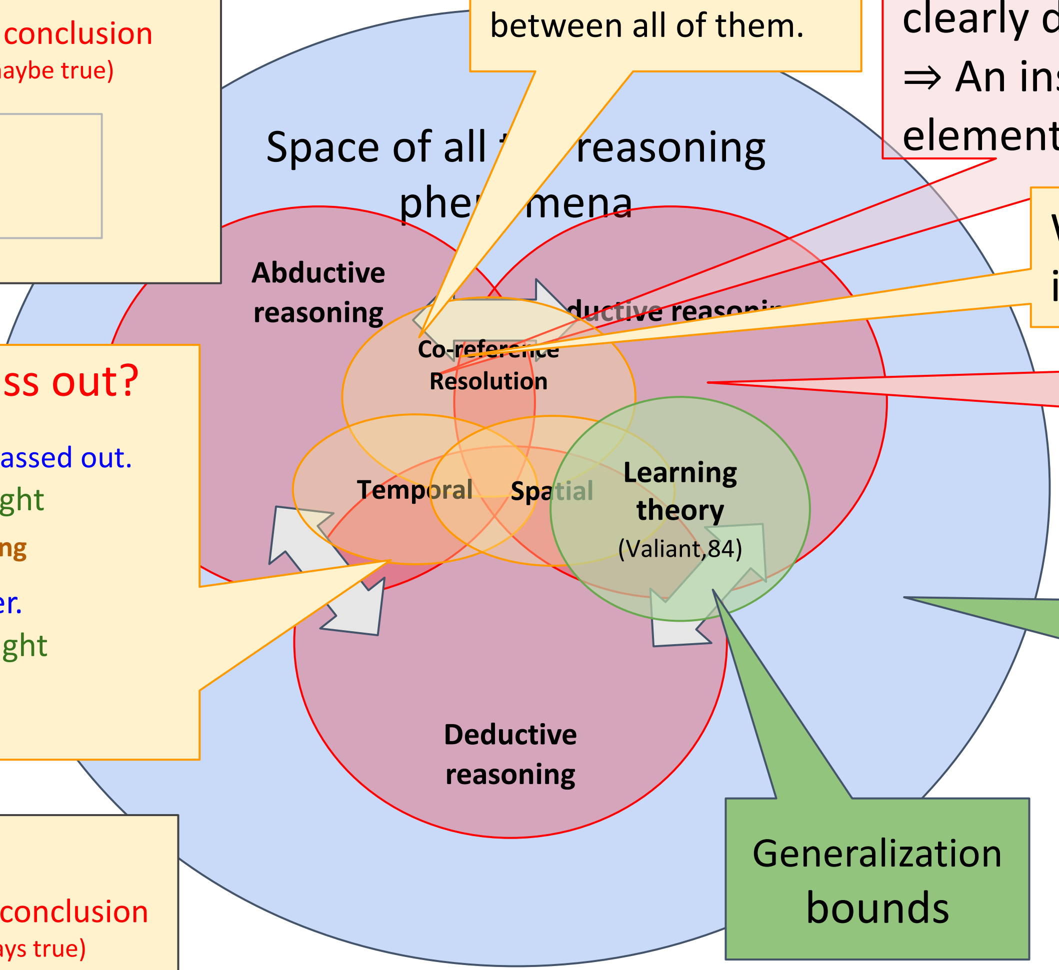
Incomplete Observations ⟹ Best conclusion (maybe true)

The grass is wet, …
- It must have rained.
- Someone has watered them

(Bayesian Nets; Fuzzy Logic; Dampster-Shafer Theory)

There is overlap between all of them.

In *language*, things are not clearly disjoint.
⇒ An instance might have elements of both phenomena.



What a linguist would interpret "reasoning"

What a logician would interpret as "reasoning"

Q: When did Jack pass out?

The sunlight hit Jack and he passed out.
Options: morning, noon, night
⇒ **Abduction: (probably) morning**

Jack passed out after dinner.
Options: morning, noon, night

⇒ **Deduction: night**

Very little understanding

Generalization bounds

Deductive reasoning

General Rule ⟹ Specific conclusion (always true)

When it rains, objects get wet.
It rained.
- The grass must be wet.

(modus ponens; modus tollens)

Flow of Ideas

$\mathbf{KB} \models \alpha$

- **Idea:** represent all your knowledge in First Order Logic (**KB**).
- **Given a query α, determine whether it holds in the KB: (KB implies α)**

- For efficiency reasons:
  - FOL (too complex to compute with) → Propositional Logic

Facts:

- Joe is married to Sue
- Bill has a brother with no children.
- Henry's friends are Bill's cousins.

(Declarative) Knowledge:

- *Ancestor* is the transitive closure of *parent*.
- *Brother* is *sibling* restricted to males
- *Favourite-cousin* is a special type of *cousin*.

Representation:

$$\forall x\, \mathrm{Friend}(\mathrm{henry}, x) \equiv \mathrm{Cousin}(\mathrm{bill}, x)$$

- **Problem I:** complexity of inference.
  - Key solution: relax expressivity.
- **(but of, course, there were many other problems** – incomplete knowledge, uncertainty)
  - E.g., what if the knowledge is not given, but rather learned?

Proof Systems

- **Given a query α, determine whether it holds in the KB: (KB implies α)**
  - Assume that KB is a collection of propositional rules: **p→q** ; this is equivalent to: **¬p**∨**q**≡ $\boldsymbol{T}$ (a tautology)
    - **p** itself can be a conjunction of propositions;
    - **q** can be a disjunction of propositions (if it a conjunction, we'll split to multiple rules.)
  - Then the KB is a conjunction of disjunctions: **a CNF**
  - Answering KB ⊨ α is equivalent to solving satisfiability for KB ∧ ¬α
    - Determining that KB ∧ ¬α has no satisfying assignments.
    - There is a lot of algorithmic proof theory to develop, under some conditions, efficient algorithms for KB ⊨ α
      - E.g., if all the rule in KB are Horn rules (monotone antecedent, a single head proposition) there is an efficient algorithm.

Proof Systems

- **Given a query α, determine whether it holds in the KB: (KB implies α)**

- But, exact reasoning could be too hard.
- And, what if KB is only approximate?

  - Model theory may makes more sense here.
    - KB ⊨ α means that all the assignments that satisfy KB also satisfy α.
    - Of course, there are too many assignments…
  - PAC semantics: what if you “sample” KB.
    - See Learning to Reason,(Khardon & Roth 96); an approach that is independent of the size of KB
    - This algorithm is complete, but not sound.
      - If KB ⊨ α it never errs. Otherwise, it may not find a counter examples.
  - It is also possible, under some conditions, to develop **exact** Learning to Reason
    - Under some assumptions on the type of queries, it is possible to find a polynomial size set of examples in KB such that is sufficient to test the query on these.

Many Other Representations

- Limited Forms of FOL
- Relational Databases: Course(csc248) Dept(csc248,ComputerScience) Enrollment(csc248,42)
  Course(mat100) Dept(mat100,Mathematics)
  - And the hope is that you can address questions such as:

    How many courses are offered by the Computer Science Department?

  - Many other representations were developed, some along with inference systems.
    - Logic Programs (Prolog): a collection of Horn sentences

      $\forall x_1 \cdots x_n [P_1 \wedge \cdots \wedge P_m \supset P_{m+1}]$ where $m \geq 0$ and each P_i is atomic

    - For example:

```
parent(bill,mary).
parent(bill,sam).
mother(X,Y) :- parent(X,Y), female(Y)
female(mary).
```

    - Now I can infer who is the Mother of Bill (if I execute the program)

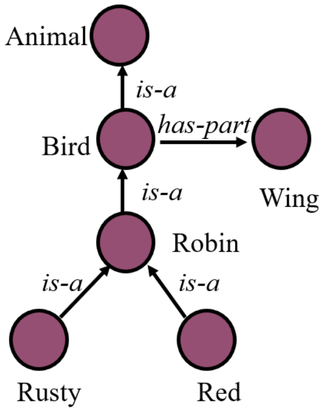
Relational Models

- Knowledge:
  - □ Actor(a) ⇒ ¬Director(a)
  - □ Director(a) ⇒ ¬WorkedFor(a,b)
  - □ InMovie(m,a) ∧ WorkedFor(a,b) ⇒ InMovie(m,b)
- Input:
  - □ Actor(Brando), Actor(Cruise), Director(Coppola),
  - □ WorkedFor(Brando, Coppola), etc.
- Query:
  - □ is (InMovie(GodFather, Brando)) ?
  - □ is (what is the probability that: Pr(InMovie(GodFather, Brando)) = ?
- Abductive version:
  - □ What is the most likely table for InMovie?

Semantic Networks

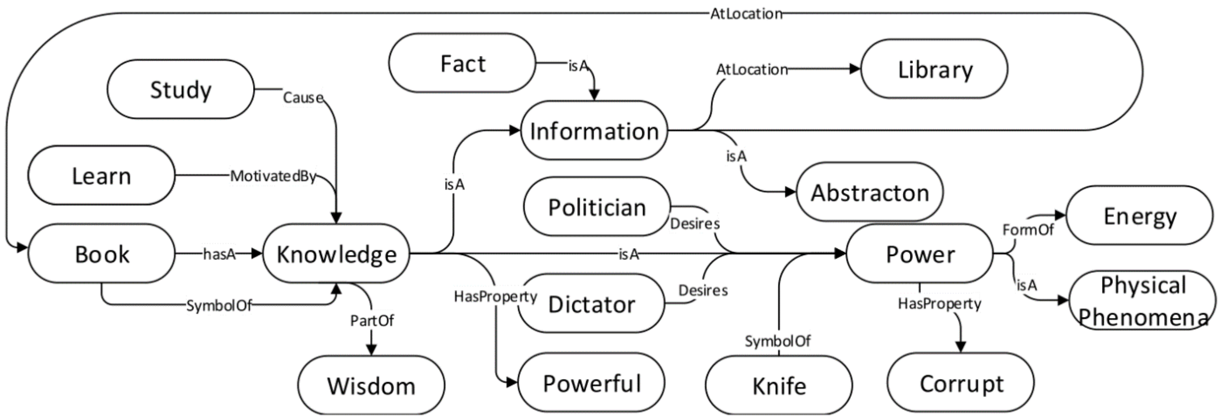
- Semantic Networks: allows the use of more expressive predicates, and more "intuitive inference".
  - People talked about inference as a form of "spreading activation"
    - A graph of labeled nodes and labeled, directed arcs
    - Arcs define relationships that hold between objects denoted by the nodes.

Link Type	Semantic s	Example
$A \xrightarrow{Subset} B$	$A \subset B$	$Cats \subset Mammals$
$A \xrightarrow{Member} B$	$A \in B$	$Bill \in Cats$
$A \xrightarrow{R} B$	$R(A, B)$	$Bill \xrightarrow{Age} 12$
$A \xrightarrow{\boxed{R}} B$	$\forall x, x \in A \Rightarrow R(x, B)$	$Bird \xrightarrow{\boxed{legs}} 12$
$A \xrightarrow{\boxed{\boxed{R}}} B$	$\forall x\ \exists y, x \in A \Rightarrow y \in B \wedge R(x, B)$	$Birds \xrightarrow{\boxed{\boxed{Parent}}} Birds$



More Networks

- This led to two directions:
- **(1) Concept nets:**
  - ☐ Based on Open Mind Common Sense (OMCS)
  - ☐ Intended to serve as a large commonsense knowledge base
  - ☐ Built on contributions of many people across the Web.



More Networks

- (2) Formalization efforts:
  - These networks were formalized in terms of **Description Logics**, and then elaborated into **Frame Description Forms**.
  - **Frames** were used to describe types and their attributes: values, Restrictions, attached procedures (how an attribute should be used).

```
(Student
    with a dept is computer-science and
    with ≥ 3 enrolled-course is a
      (Graduate-Course
          with a dept is a Engineering-Department))
```

  - Eventually, this led to theories of **Frames** (Minsky), and **Scripts** (Schank)
    - There are beginning to be influential again, where people think more about **Events**

More Networks

- More generally, these languages had expressive grammars:

$$
\begin{aligned}
\langle type \rangle ::= {} & \langle atom \rangle \\
& \mid (\texttt{AND}\ \langle type_1 \rangle \ldots \langle type_n \rangle) \\
& \mid (\texttt{ALL}\ \langle attribute \rangle\ \langle type \rangle) \\
& \mid (\texttt{SOME}\ \langle attribute \rangle)
\end{aligned}
$$

$$
\begin{aligned}
\langle attribute \rangle ::= {} & \langle atom \rangle \\
& \mid (\texttt{RESTR}\ \langle attribute \rangle\ \langle type \rangle)
\end{aligned}
$$

  - Example: The set of all people the all their male friends are doctors with some specialty.

(AND person (ALL (RESTR friend male) (AND doctor (SOME specialty)))).

- And it came with inference algorithms – **subsumption,** and was extremely influential – all systems built in the 80-ith and later, were built on these languages.
  - It was also influential in areas such as Feature Extraction for machine learning, and theories of grammar.

Relational Models

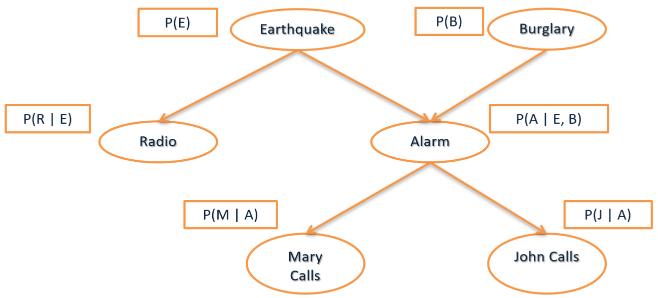
- Knowledge:
  - Actor(a) ⇒ ¬Director(a)
  - Director(a) ⇒ ¬WorkedFor(a,b)
  - InMovie(m,a) ∧ WorkedFor(a,b) ⇒ InMovie(m,b)
- Input:
  - Actor(Brando), Actor(Cruise), Director(Coppola),
  - WorkedFor(Brando, Coppola), etc.
- Query:
  - is (InMovie(GodFather, Brando)) ?
  - is (what is the probability that: Pr(InMovie(GodFather, Brando)) = ?
- Abductive version:
  - What is the most likely table for InMovie?

Probabilities

- In parallel to the progress on the logical representations, people argued that we need to deal with uncertainly, and need to move to **probabilistic representations.**
- Progress here proceeded in two camps
  - (Propositional) representation of distributions
    - **Bayesian Networks** (Pearl 1988)
  - Probabilistic extensions of the FOL/Prolog representations. (Haddawy 1993)
    - **Problog**
    - Markov Logic Network

- Two important comments:
  - The latter direction is presented today as fusing probabilities with declarative (logical) knowledge. This, in fact, was studies much earlier (in the 60—ies), but without practical implementations.
  - Fusing Probabilities with Declarative information is different from fusing Learning with Declarative Information. In fact, none of the bullets above came with a native approach for **learning**.
  - Fusing learning with declarative knowledge came later in the context of Structured Learning, e.g., ILP Formulations, Roth & Yih 2004, and following works.

Bayes Nets



- Nodes are random variables
- Edges represent causal influences
- Each node is associated with a conditional Probability distribution
- **Computational Problems (**Inference):
  - **Computing the probability of an event:**
  - Given structure and parameters
  - Given an observation E, what is the probability of assignment Y?
  - **P(R=off, A=off | E=e) =?** (E, Y are sets of instantiated variables)
- **Most likely explanation (Maximum A Posteriori assignment, MAP, MPE)**
  - Given structure and parameters
  - Given an observation E, what is the most likely assignment to Y?
  - Argmax$_y$ **P(Y=y | E=e) (Say, Y = (R, A))**
  - (E, Y are sets of instantiated variables)

Probabilistic Relational Representations

- Representation of distributions over relations, as opposed to propositional variables.
- Ability to build programs that do not only encode complex interactions between variables but also express inherent uncertainty.

- **Inference:** Becoming much harder. For the most part, done by **propositionalizing** relational representations (that is, substitution of all domain variables, and blowing up the representations to get a propositional BN).
- But, there are other ways, e.g., **lifted inference**.

```
0.3::stress(X) :- person(X).
0.2::influences(X,Y) :- person(X), person(Y).

smokes(X) :- stress(X).
smokes(X) :- friend(X,Y), influences(Y,X), smokes(Y).

0.4::asthma(X) :- smokes(X).

person(angelika).
person(joris).
person(jonas).
person(dimitar).

friend(joris,jonas).
friend(joris,angelika).
friend(joris,dimitar).
friend(angelika,jonas).
```