Texts as Knowledge Bases

Christopher Manning

Joint work with Gabor Angeli and Danqi Chen

Stanford NLP Group

@chrmanning · @stanfordnlp

AKBC 2016

Sean Goldberg
@SeanGoldbergCS

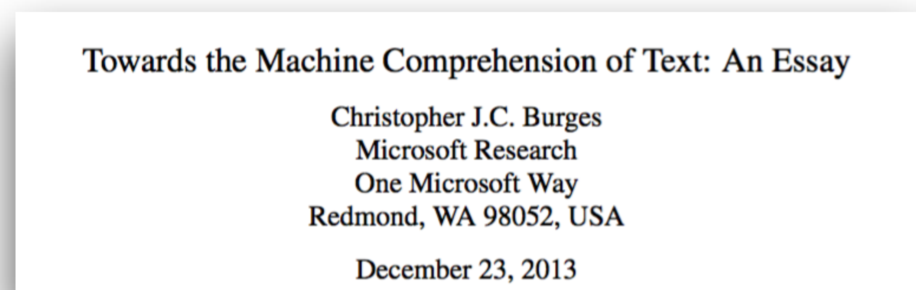
Knowledge Acquisition sess. @ #naacl2016 still stuck in using clean,fixed ontology KBs. Bigger problem is incorp. info from open, noisy KBs.

6/14/16, 11:58 AM

1 LIKE

Machine Comprehension = Machine has an Augmented Knowledge Base

"A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."



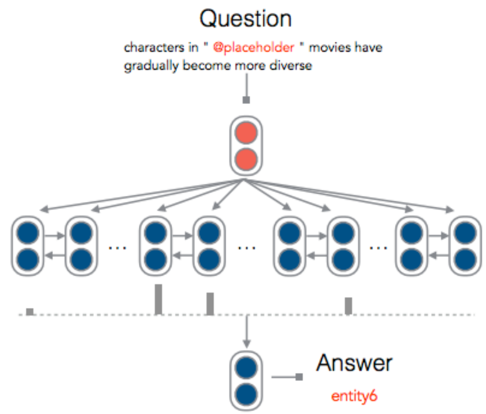
Towards the Machine Comprehension of Text: An Essay

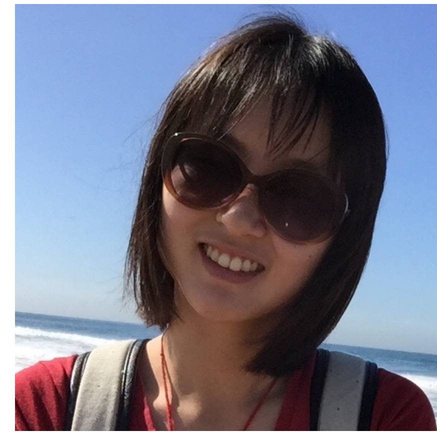
Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

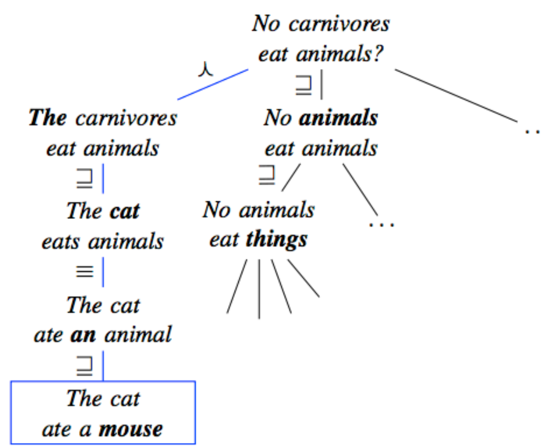
Two case studies … previews of ACL 2016

How far do current deep learning reading comprehension systems go in achieving Chris Burges's goal?





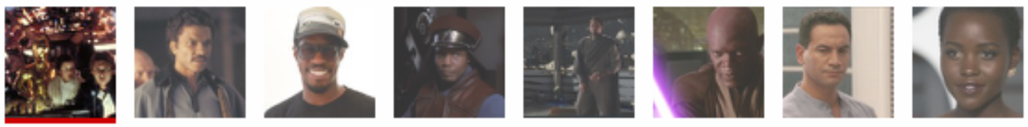
How can we use natural logic and shallow reasoning to better treat texts as a knowledge base?



DeepMind RC dataset

[Hermann et al. 2015]

CNN Entertainment » 'Star Wars' universe gets its first gay character



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in _______ movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

DeepMind RC dataset

Large data set

Real language

Good for DL training!

"Artificial" pre-processing (coref, anonymization)

How hard?

Is it a good task?

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question	Answer
characters in " @placeholder " movies have gradually become more diverse	@entity6

Results on DeepMind RC when we began

[Hermann et al. 2015; Hill et al. 2016]

System	CNN Dev	CNN Test	Daily Mail Dev	Daily Mail Test
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Attentive Reader	61.6	63.0	70.5	69.0
Impatient Reader	61.8	63.8	69.0	68.0
MemNN window memory	58.0	60.6		
MemNN window + self sup	63.4	66.8		
MemNN win, ss, ens, no-c	66.2	69.4		

Frame semantics or simple syntax?

Frame-semantic parsing attempts to identify predicates and their semantic arguments – should be good for question answering!

Hermann et al. use a "state-of-the-art frame-semantic parser" – Google version of [Das et al. 2013, Hermann et al. 2014]

But frame semantic systems have coverage problems, not representing pertinent relations not mapped onto verbal frames

How about a good old feature-based system, using a syntactic dependency parser?

System I: Standard Entity-Centric Classifier

[Chen, Bolton, & Manning, ACL 2016]

- Build a symbolic feature vector for each entity: $f_{p,q}(e)$
- The goal is to learn feature weights such that the correct answer ranks higher than the other entities
- Train logistic regression and MART classifier (boosted decision trees – these do better and are reported)
    - Whether *e* is in the passage
    - Whether *e* is in the question
    - Frequency of *e* in passage
    - First position of *e* in passage
    - n-gram exact match (features for matching L/R 1/2 words)
    - Word distance of question words in passage
    - Whether *e* co-occurs with *q* verb or another entity
    - Syntactic dependency parse triple match around *e*

Competent (traditional) statistical NLP …

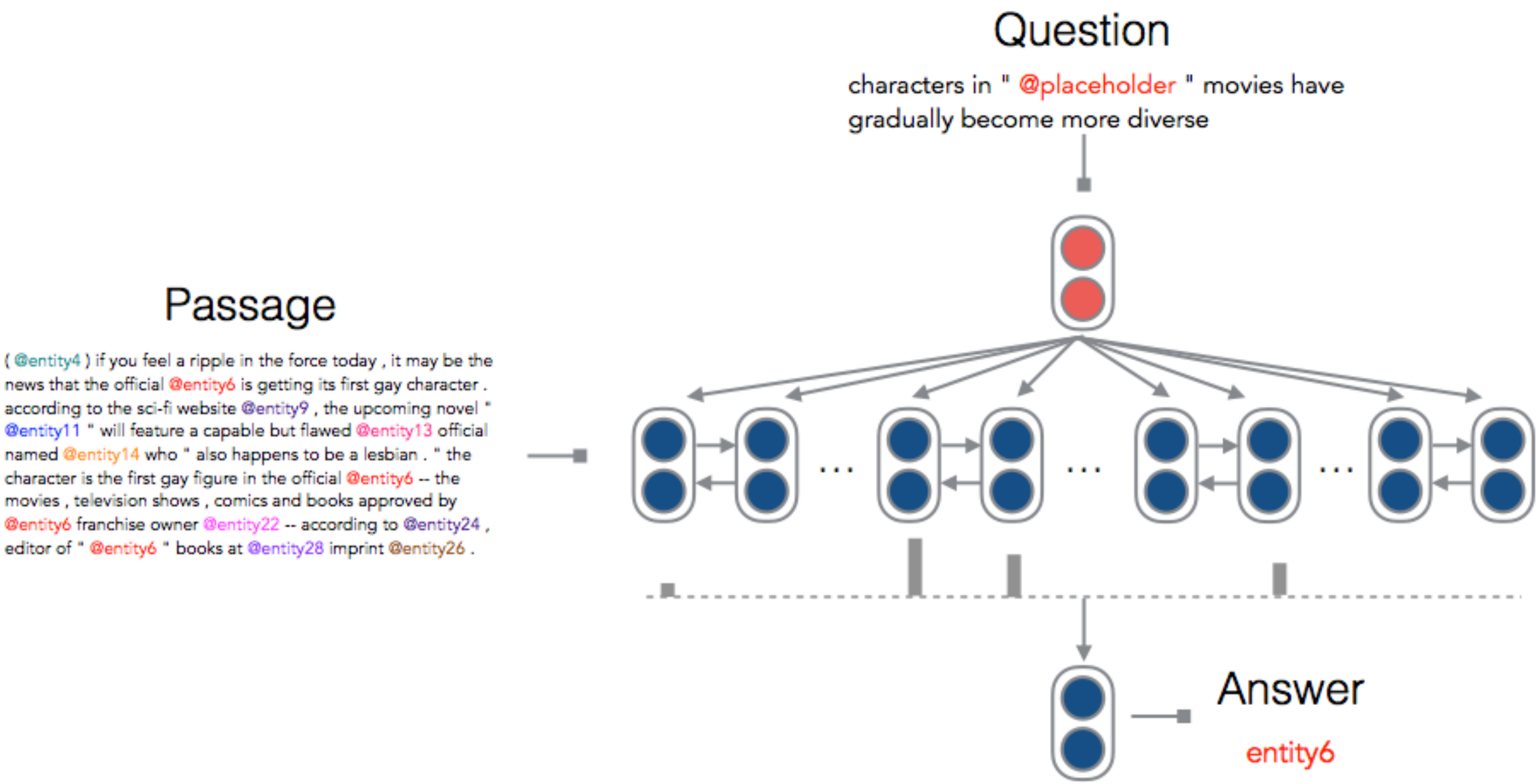
System	CNN Dev	CNN Test	Daily Mail Dev	Daily Mail Test
Frame-semantic model	36.3	40.2	35.5	35.5
Impatient Reader	61.8	63.8	69.0	68.0
Competent statistical NLP	**67.1**	**67.9**	**69.1**	**68.3**
MemNN window + self sup	63.4	66.8		
MemNN win, ss, ens, no-c	66.2	**69.4**		

Ablating individual features

Features	Accuracy
Full model	67.1
– whether *e* is in the passage	67.1
– whether *e* is in the question	67.0
– frequency of *e*	**63.7**
– position of *e*	65.9
– *n*-gram match	**60.5**
– word distance	65.4
– sentence co-occurrence	66.0
– dependency parse match	65.6

System II: End-to-End Neural Network

[Chen, Bolton, & Manning, ACL 2016]



System II: End-to-End Neural Network

No magic at all; we make our model as simple as possible

- Learned word embeddings feed into
- Bi-directional shallow LSTMs for passage and question
- Question representation used for soft attention over passage with **simple bilinear attention function**

$$\alpha_i = \operatorname{softmax}_i \left(\mathbf{q}^\top \mathbf{W}_s \tilde{\mathbf{p}}_i\right)$$
$$\mathbf{o} = \sum_i \alpha_i \tilde{\mathbf{p}}_i$$

- A final softmax layer predicts the answer entity
- SGD, dropout (0.2), batch size = 32, hidden size = 128, …

Competent new-fangled NLP …

System	CNN Dev	CNN Test	DM Dev	DM Test
Impatient Reader	61.8	63.8	69.0	68.0
Competent statistical NLP	67.1	67.9	69.1	68.3
Our LSTM with attention	**72.4**	**72.4**	**76.9**	**75.8**
MemNN window + self sup	63.4	66.8		
MemNN win, ss, ensem, no-c	66.2	69.4		

Differences:

Simple bilinear attention [Luong, Pham, & Manning 2015]

Hermann et al. had an extra, unnecessary layer joining *o* and *q*

We predict among entities, not all words (but doesn't make a difference)

Maybe we're better at tuning neural nets? Been doing it for a while.

Our Results

We are quite happy with the numbers 😊

[and, BTW, several other people have now gotten similar numbers]

… but what do they really mean?

- What level of language understanding is needed?
- What have the models actually learned?

Data Analysis

A breakdown of the examples

Exact match

Sentence-level paraphrasing / textual entailment

Partial clue

Multiple sentences

Coreference errors

Ambiguous or too hard

Category	Question	Passage
Exact Match	*it 's clear @entity0 is leaning toward* **@placeholder** , says an expert who monitors @entity0	. . . @entity116 , who follows @entity0 's operations and propaganda closely , recently told @entity3 , *it 's clear @entity0 is leaning toward* **@entity60** in terms of doctrine , ideology and an emphasis on holding territory after operations . . . .
Paraphrase	**@placeholder** says he understands why @entity0 wo n't play at his tournament	. . . @entity0 called me personally to let me know that he would n't be playing here at @entity23 , " **@entity3** said on his @entity21 event 's website . . . .
Partial clue	a tv movie based on @entity2 's book **@placeholder** casts a @entity76 actor as @entity5	. . . to @entity12 @entity2 professed that his **@entity11** is not a religious book . . . .
Multiple sent.	he 's doing a his - and - her duet all by himself , @entity6 said of **@placeholder**	. . . we got some groundbreaking performances , here too , tonight , @entity6 said . we got **@entity17** , who will be doing some musical performances . he 's doing a his - and - her duet all by himself . . . .
Coref. Error	rapper **@placeholder** " disgusted , " cancels upcoming show for @entity280	. . . with hip - hop star **@entity246** saying on @entity247 that he was canceling an upcoming show for the @entity249 . . . . (but @entity249 = @entity280 = SAEs)
Hard	pilot error and snow were reasons stated for **@placeholder** plane crash	. . . a small aircraft carrying **@entity5** , @entity6 and @entity7 the @entity12 @entity3 crashed a few miles from @entity9 , near @entity10 , @entity11 . . . .

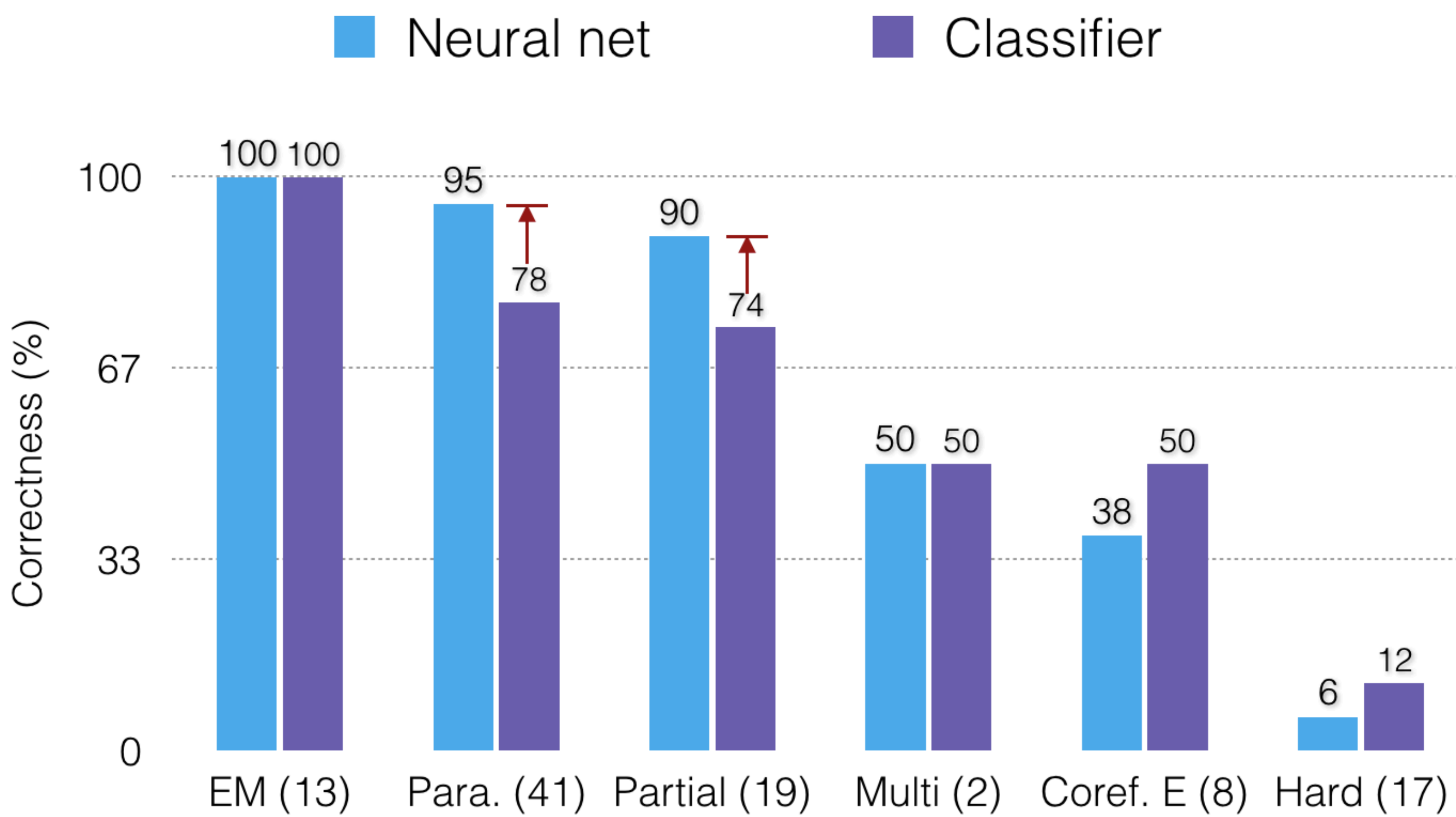
Data Analysis

No.	Category	(%)
1	Exact match	13
2	Paraphrasing	41
3	Partial clue	19
4	Multiple sentences	2
5	Coreference errors	8
6	Ambiguous / hard	17

- **25%**: coreference errors + hard cases
- Only **2%** require multiple sentences

Ours: Classifier	67.1	67.9	69.1	68.3
Ours: Neural net	**72.4**	**72.4**	**76.9**	**75.8**

Data Analysis



Discussion

- The DeepMind RC data is quite noisy
- The required reasoning and inference level is quite limited
- There isn't much room left for improvement
- However, the scale and ease of data production is appealing
- Can we make use of this data in solving more realistic RC tasks?

- Neural networks are *great* for learning semantic matches across lexical variation or paraphrasing!
- LSTMs with (simple bilinear) attention are *great!*
- Not yet proven whether NNs can do more challenging RC tasks

AI2 4th Grade Science Question Answering

[Angeli, Nayak, & Manning, ACL 2016]

Our "knowledge":

Ovaries are the female part of the flower, which produces eggs that are needed for making seeds.

The question:

Which part of a plant produces the seeds?

The answer choices:

the flower *the leaves* *the stem* *the roots*

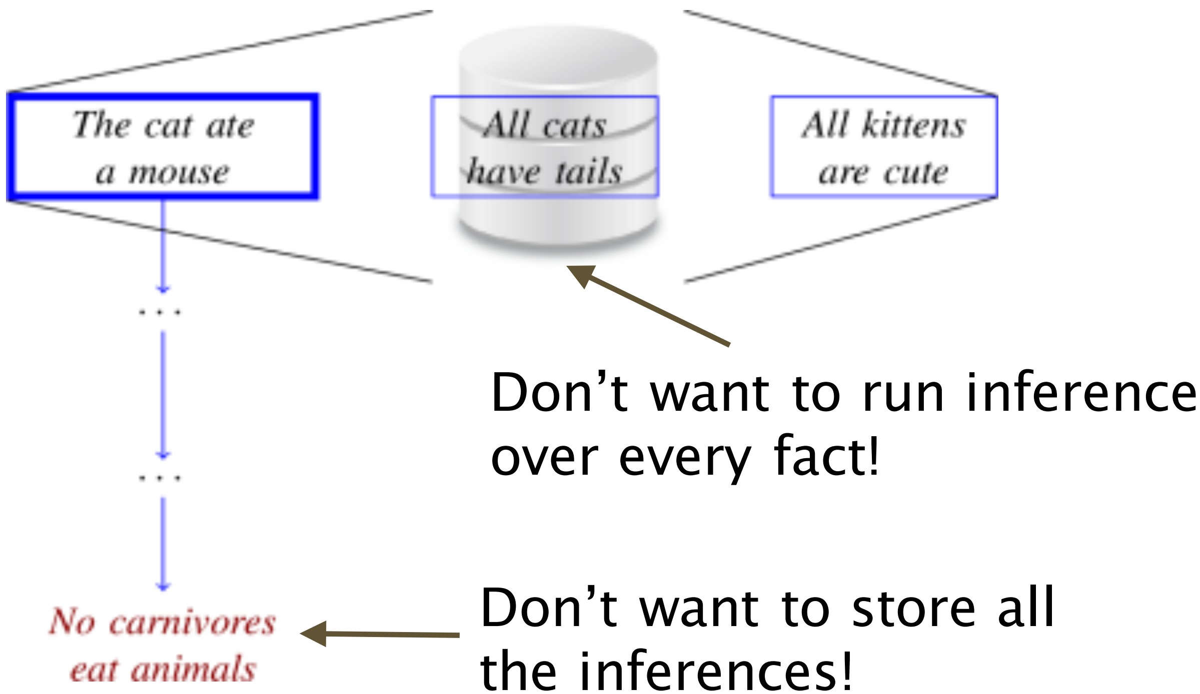
How can we represent and reason with broad-coverage knowledge?

1. Rigid-schema knowledge bases with well-defined logical inference
2. Open-domain knowledge bases (Open IE) – no clear ontology or inference [Etzioni et al. 2007ff]
3. Human language text KB – No rigid schema, but with "Natural logic" can do formal inference over human language text

Text as Knowledge Base

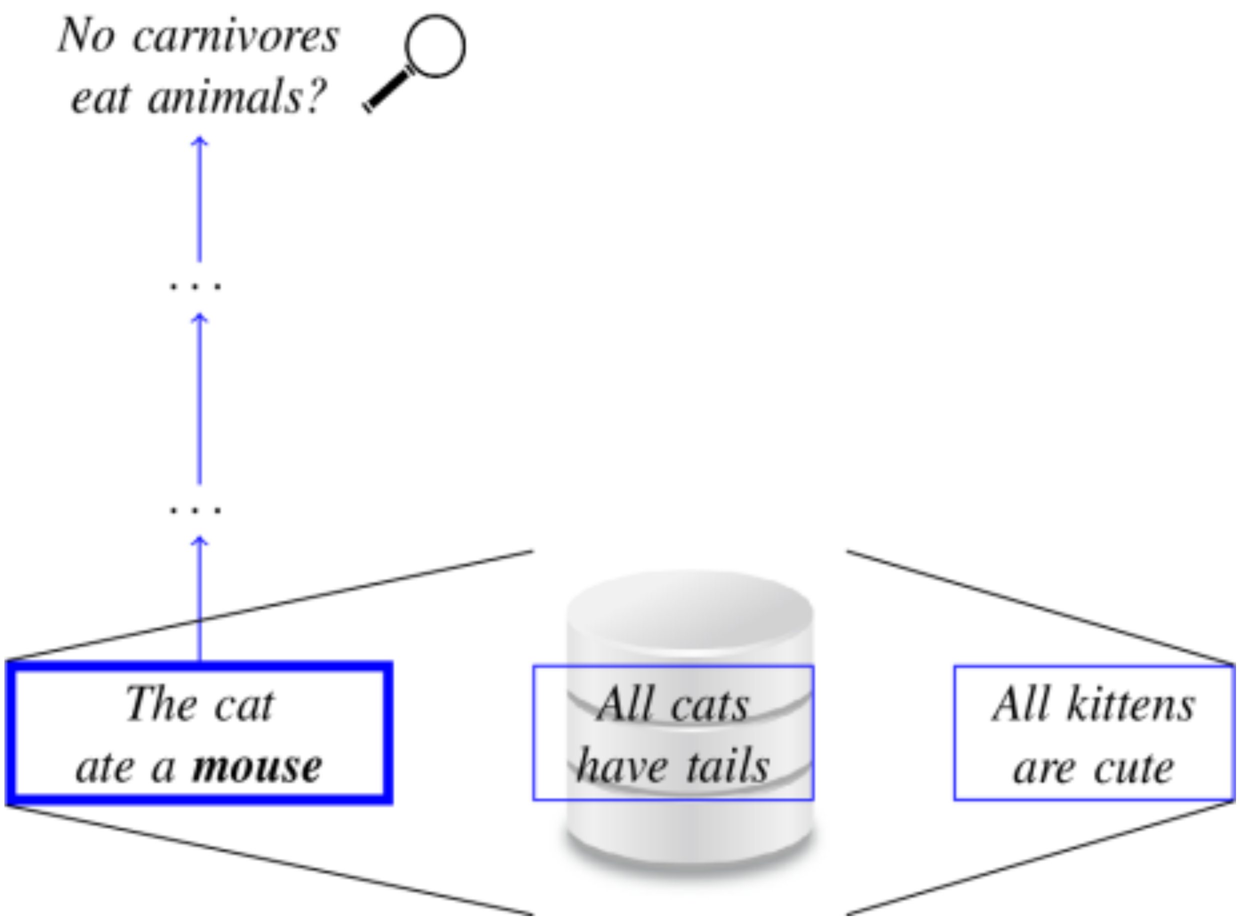
Storing knowledge as text is easy!

Doing inferences over text might be hard

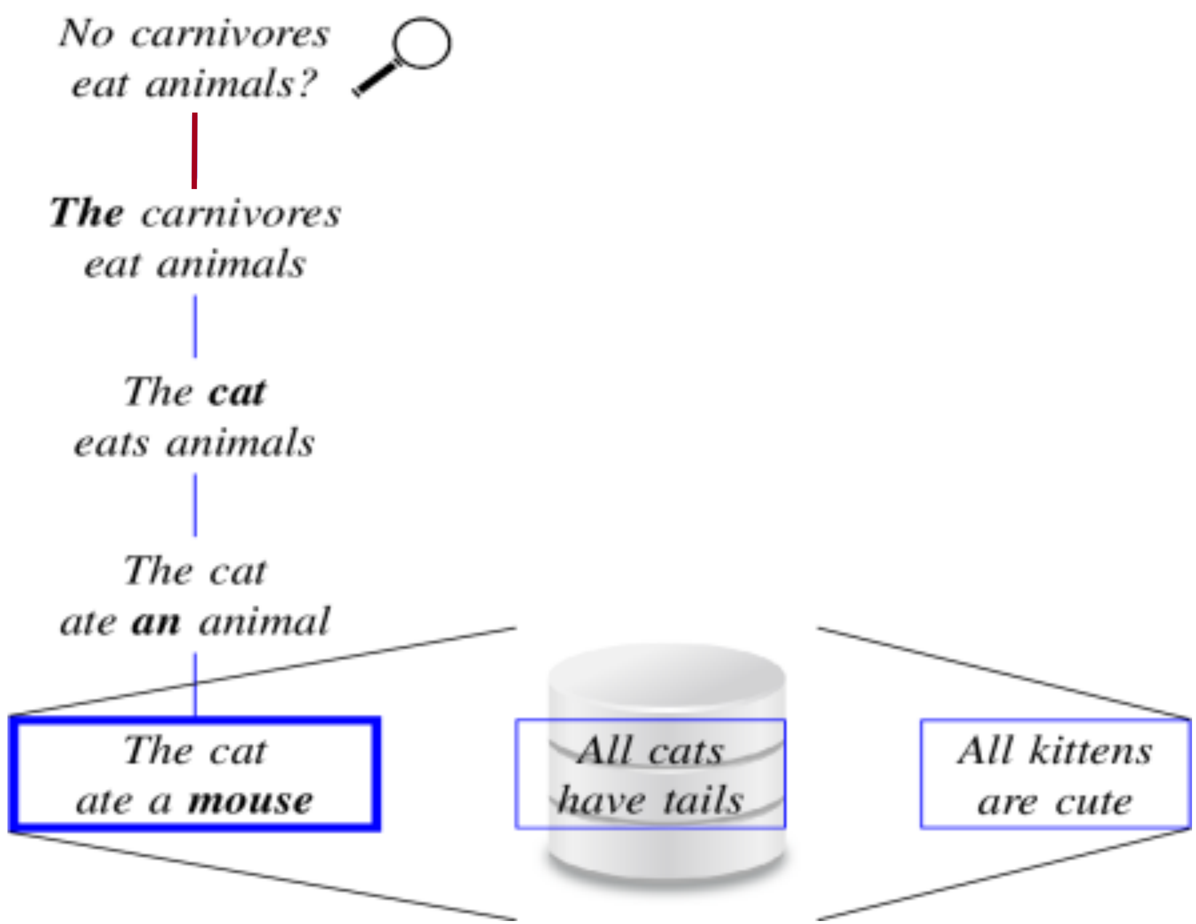


Inferences … on demand from a query…

[Angeli and Manning 2014]



… using text as the meaning representation



Natural Logic: logical inference over text

We are doing logical inference

The cat ate a mouse ⊨ ¬ *No carnivores eat animals*

We do it with natural logic

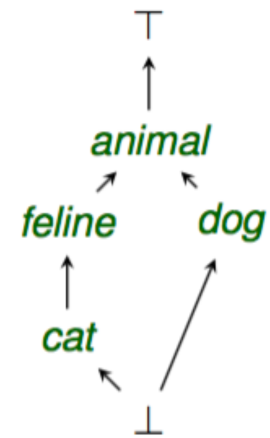
If I mutate a sentence in this way, do I preserve its truth?

Post-Deal Iran Asks if U.S. Is Still 'Great Satan,' or Something Less ⊨
A Country Asks if U.S. Is Still 'Great Satan,' or Something Less

- A sound and complete weak logic [Icard and Moss 2014]
- Expressive for common human inferences*
- "Semantic" parsing is just syntactic parsing
- Tractable: Polynomial time entailment checking
- Plays nicely with lexical matching back-off methods

#1. Common sense reasoning

Polarity in Natural Logic



We order phrases in *partial orders* (not just is-a-kind-of, can also do geographical containment, etc.)

Polarity is the direction a phrase can move in this order

Example inferences

Quantifiers determine the *polarity* of phrases

Valid mutations consider polarity

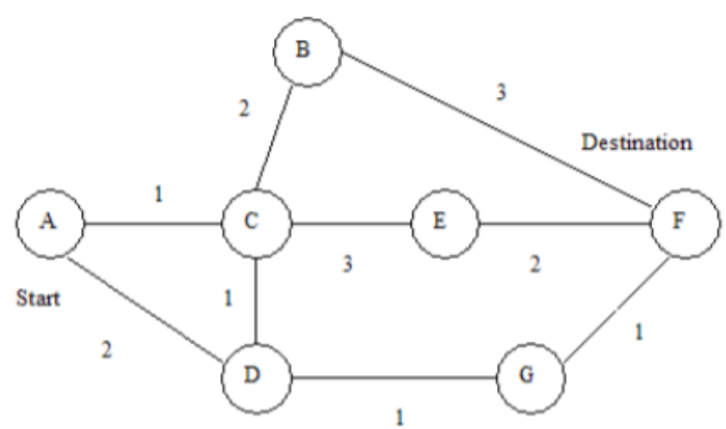
Successful toy inference:

All cats eat mice $\models$ *All house cats consume rodents*

"Soft" Natural Logic

We also want to make likely (but not certain) inferences

- Same motivation as Markov logic, probabilistic soft logic, etc.
- Each mutation *edge template* has a cost $\theta \geq 0$
- Cost of an edge is $\theta_i \cdot f_i$
- Cost of a path is $\boldsymbol{\theta} \cdot \mathbf{f}$
- Can learn parameters $\boldsymbol{\theta}$
- **Inference is then graph search**



#2. Dealing with real, long sentences

Natural logic works with facts like these in the knowledge base:

Obama was born in Hawaii

But real-world sentences are complex:

***Born in Honolulu, Hawaii, Obama** is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review.*

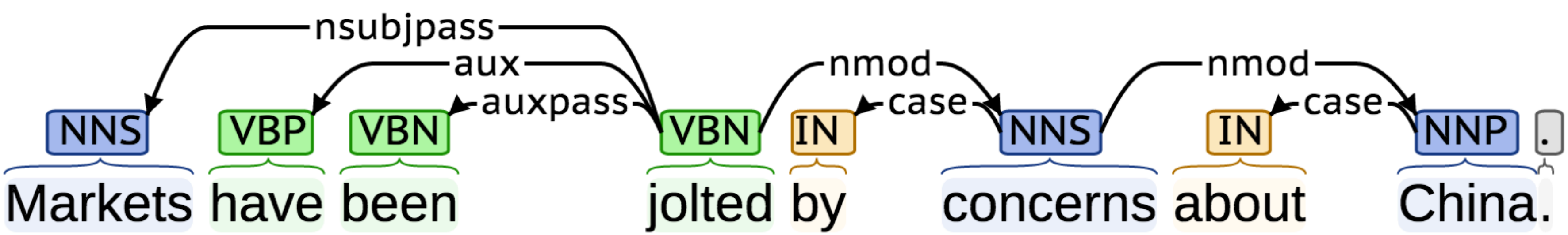
Approach:

1. Classifier yields entailed clauses from a long sentence
2. Shorten clauses with natural logic inference

Universal Dependencies (UD)

http://universaldependencies.github.io/docs/

A single level of typed dependency syntax that gives a simple, human-friendly representation of sentence structure and meaning

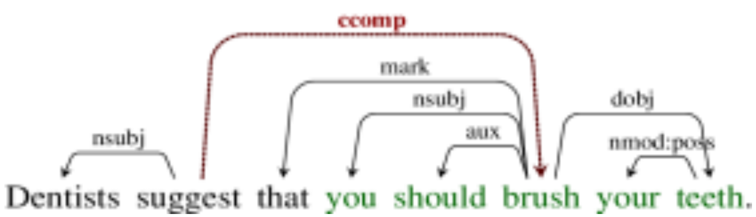


Better than a phrase-structure tree for machine interpretation – it's almost a semantic network

UD aims to be **linguistically better across languages** than earlier, common, simple NLP representations, such as CoNLL dependencies

Generation of minimal clauses

1. Classification problem: given a dependency edge, is it a clause?



2. Is it missing a controlled subject from subj/object?

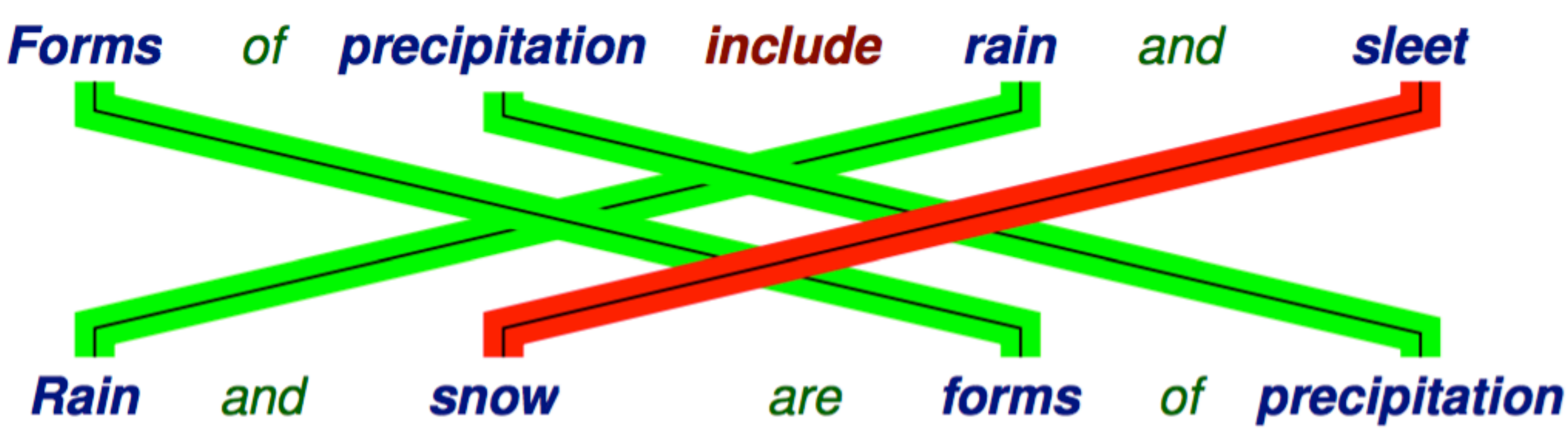
3. Shorten clauses while preserving validity!

- *All **young rabbits** drink milk ⊭ All **rabbits** drink milk*
- **OK:** *SJC, the bay area's third largest airport, is experiencing delays due to weather.*
- **Often better:** *SJC is experiencing delays.*

Using **natural logic**

#3. Add a lexical alignment classifier

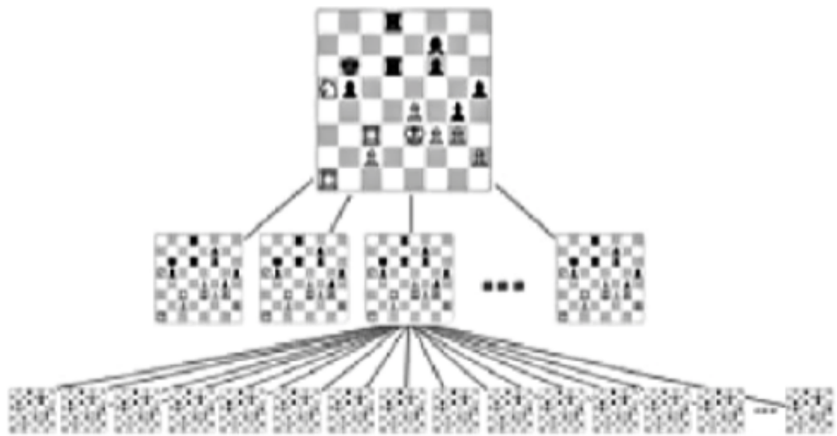
- Sometimes we can't quite make the inferences that we would like to make:



- We use a simple lexical match back-off classifier with features:
  - Matching words, mismatched words, unmatched words
  - These always work pretty well – **the** lesson of RTE evaluations

The full system

- We run our usual search over split up, shortened clauses
  - If we find a premise, great!
  - If not, we use the lexical classifier as an *evaluation function*



  - We work to do this quickly
    - Visit 1M nodes/second, don't refeaturize, just delta
    - 32 byte search states (thanks Gabor!)

Solving 4th grade science (Allen AI datasets)

Multiple choice questions from real 4th grade science exams

Which activity is an example of a good health habit?

(A) Watching television (B) Smoking cigarettes (C) Eating candy (D) Exercising every day

In our ~~corpus~~ knowledge base:

- *Plasma TV's can display up to 16 million colors ... great for watching TV ... also make a good screen.*
- *Not smoking or drinking alcohol is good for health, regardless of whether clothing is worn or not.*
- *Eating candy for diner is an example of a poor health habit.*
- *Healthy is exercising*

Solving 4th grade science (Allen AI NDMC)

System	Dev	Test
KnowBot [Hixon et al. NAACL 2015]	45	–
KnowBot (Oracle – human in loop)	57	–
IR baseline (Lucene)	49	42
NaturalLI	52	51
More data + IR baseline	62	58
More data + NaturalLI	65	61
NaturalLI +  +  (lex. classifier)	74	**67**
Aristo [Clark et al. 2016] 6 systems, even more data		**71**

Test set: New York Regents 4th Grade Science exam multiple-choice questions from AI2
Training: Basic is Barron's study guide; more data is SciText corpus from AI2. Score: % correct

Envoi

Can our knowledge base just be text?

Natural logic provides a useful, formal (weak) logic for textual inference

Natural logic is easily combinable with lexical matching methods, including neural net methods

The resulting system is useful for:

- Common-sense reasoning
- Question Answering
- Also, Open Information Extraction