CS395T: Structured Models for NLP
Lecture 26: Wrapup

Greg Durrett

Administrivia

- Project presentations next Tuesday and Thursday
- Final projects due December 15

This Lecture

- Grammar Induction
- Structure in NLP
- Ethics in NLP

Grammar Induction

Motivation

- Parsers are important for being able to do well at machine translation, information extraction, etc.

- Expensive to annotate treebanks in new domains or languages — how do we learn parsers quickly?

Grammar Induction

- Learn a PCFG through simple EM: E-step estimates posteriors over rules, M-step re-estimates grammar

1.0	S	→	$\overline{verb}$
.2	$\overline{verb}$	→	verb
.4	$\overline{verb}$	→	$\overline{noun}$ verb
.4	$\overline{verb}$	→	verb $\overline{noun}$
.5	$\overline{noun}$	→	noun
.5	$\overline{noun}$	→	$\overline{det}$ noun

Figure 5: Intended solution for the simple corpus

Carroll and Charniak (1992)

Grammar Induction

- Learn a PCFG through simple EM: E-step estimates posteriors over rules, M-step re-estimates grammar

- Gets stuck in local optima: uses the grammar to memorize sentences rather than model language and generalize

1.0	S	$\rightarrow$	$\overline{verb}$
.2	$\overline{verb}$	$\rightarrow$	verb
.4	$\overline{verb}$	$\rightarrow$	$\overline{noun}$ verb
.4	$\overline{verb}$	$\rightarrow$	verb $\overline{noun}$
.5	$\overline{noun}$	$\rightarrow$	noun
.5	$\overline{noun}$	$\rightarrow$	$\overline{det}$ noun

Figure 5: Intended solution for the simple corpus

1.0	S	$\rightarrow$	$\overline{verb}$
.2	$\overline{verb}$	$\rightarrow$	verb
.2	$\overline{verb}$	$\rightarrow$	$\overline{noun}$ verb
.2	$\overline{verb}$	$\rightarrow$	verb $\overline{noun}$
.2	$\overline{verb}$	$\rightarrow$	$\overline{det}$ $\overline{noun}$ verb
.2	$\overline{verb}$	$\rightarrow$	verb $\overline{det}$ $\overline{noun}$

Figure 6: Solution actually found for the simple corpus

Carroll and Charniak (1992)

Dependency Model with Valence

NN	VBD	TO	NN
I	ran	to	school

- Simple generative model of dependencies:

$$P(\text{ran}|\text{root})$$

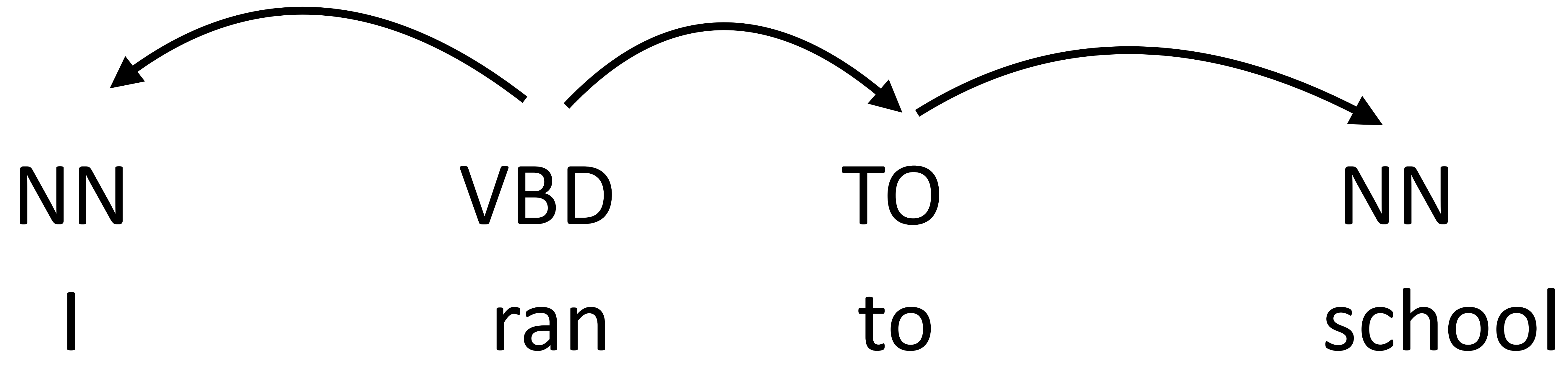
$$\cdot\, P(\text{to}|\text{ran}, \text{dir}{=}\text{right})P(\text{STOP}|\text{ran}, \text{dir}{=}\text{right}, \text{adj}{=}\text{False})$$
- generate right children

$$\cdot\, P(\text{I}|\text{ran}, \text{dir}{=}\text{left})P(\text{STOP}|\text{ran}, \text{dir}{=}\text{left}, \text{adj}{=}\text{False})$$
- generate left children

$$\cdot\, P(\text{school}|\text{to}, \text{dir}{=}\text{right})\ldots$$
- Recurse down the tree

Klein and Manning (2003)

Dependency Model with Valence



valence model

$$P(D(h)) = \prod_{dir \in \{l,r\}} \prod_{a \in deps_D(h,dir)} P_{\text{STOP}}(\neg \text{STOP} \mid h, \boxed{dir, adj}) \, P_{\text{CHOOSE}}(a \mid h, dir) P(D(a))$$

- Explicitly choose not to stop at each step

$$P_{\text{STOP}}(\text{STOP} \mid h, \boxed{dir, adj})$$

- Stop once children are generated

Klein and Manning (2003)

Dependency Model with Valence

- Use EM to learn parameters (E-step: CKY, M-step: count + normalize)

Model	UP	UR	UF_1	Dir	Undir
English (WSJ10 – 7422 Sentences)					
LBRANCH/RHEAD	25.6	32.6	28.7	33.6	56.7
RANDOM	31.0	39.4	34.7	30.1	45.6
RBRANCH/LHEAD	55.1	70.0	61.7	24.0	55.9
DMV	46.6	59.2	52.1	43.2	62.7
CCM	64.2	81.6	71.9	23.8	43.3
DMV+CCM (POS)	**69.3**	**88.0**	**77.6**	**47.5**	**64.5**

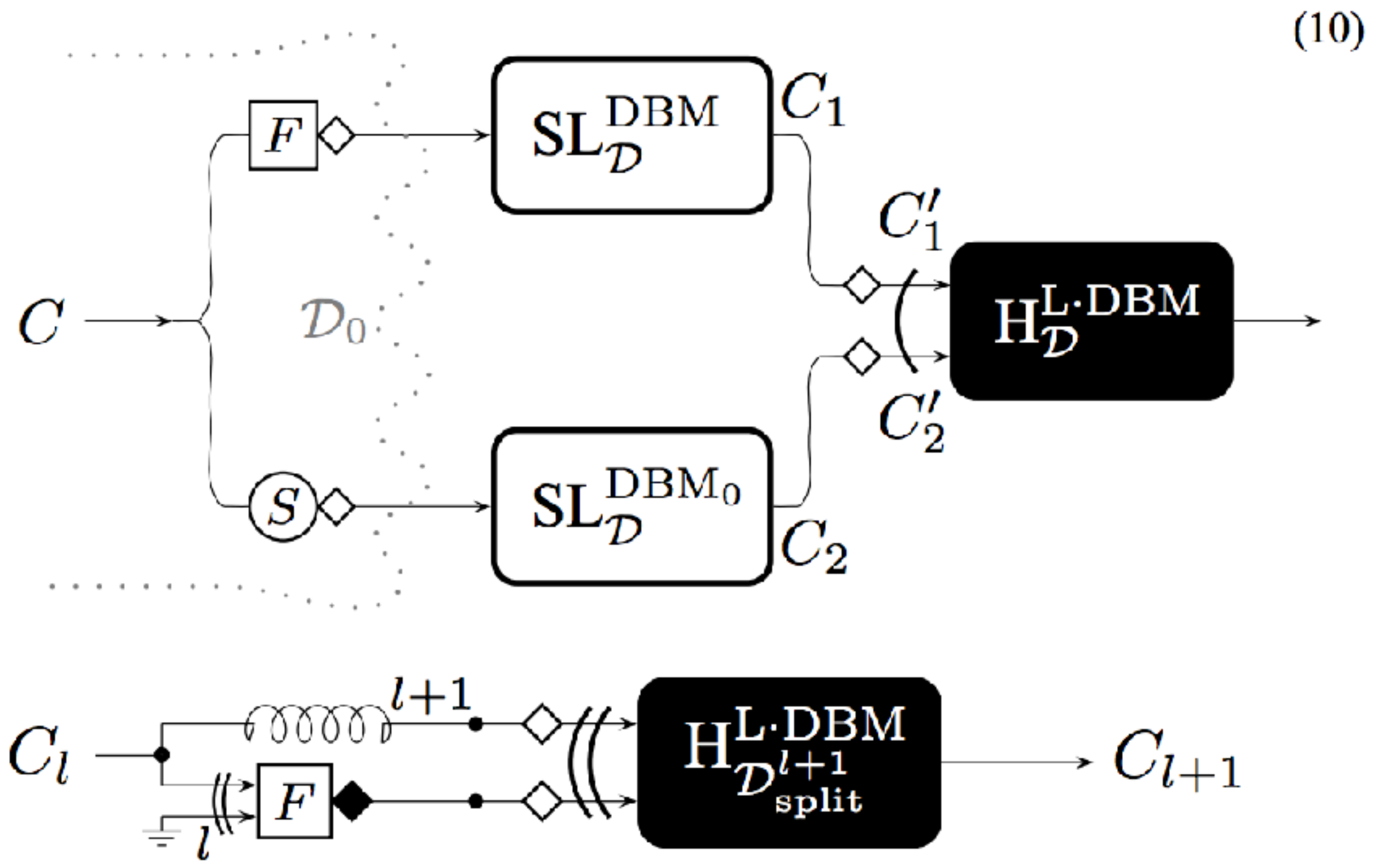
- Sentences of length 10 — incredibly easy! (supervised models >98%)
- Uses gold part-of-speech tags

Klein and Manning (2003)

Crazy DMV

- Take a learned grammar, make two copies with perturbed inputs or perturbed parameters, run EM again

- ~5 different ways like this of systematically getting better grammars, combine all of these in a big crazy system



Spitkovsky et al. (2013)

Crazy DMV

- Sentence of length 40: much harder and more real task
- Supervised parsers: 95+ UAS on this task

System		F_1	P	R
Binary-Branching	Upper Bound	85.7		
Left-Branching	Baseline	12.0		
CCM	(Klein and Manning, 2002)	33.7		
Right-Branching	Baseline	40.7		
F-CCM	(Huang et al., 2012)	45.1		
HMM	(Ponvert et al., 2011)	46.3		
LLCCM	(Golland et al., 2012)	47.6		
CCL	(Seginer, 2007)	52.8	54.6	51.1
PRLG	(Ponvert et al., 2011)	**54.6**	**60.4**	49.8
CS	**System Combination**	**54.2**	55.6	**52.8**
Supervised DBM	Skyline	59.3	65.7	54.1
Dependency-Based	Upper Bound	87.2	100	77.3

Spitkovsky et al. (2013)

What about small amounts of supervision?

Supervision from other languages: train a parser over universal POS tags on a different language entirely

	DELEX
DA	41.3
DE	58.5
EL	57.9
ES	64.2
IT	65.9
NL	57.0
PT	75.4
SV	64.5
AVG	60.6

Train on 100 trees, also use a bilingual dictionary

	DELEX	DELEX+PROJ
DA	67.2	69.5
DE	72.9	73.9
EL	70.8	72.9
ES	72.5	73.0
IT	73.3	75.4
NL	63.0	65.8
PT	78.1	79.5
SV	76.4	78.1
AVG	71.8	73.5

McDonald et al. (2011), Durrett et al. (2012)

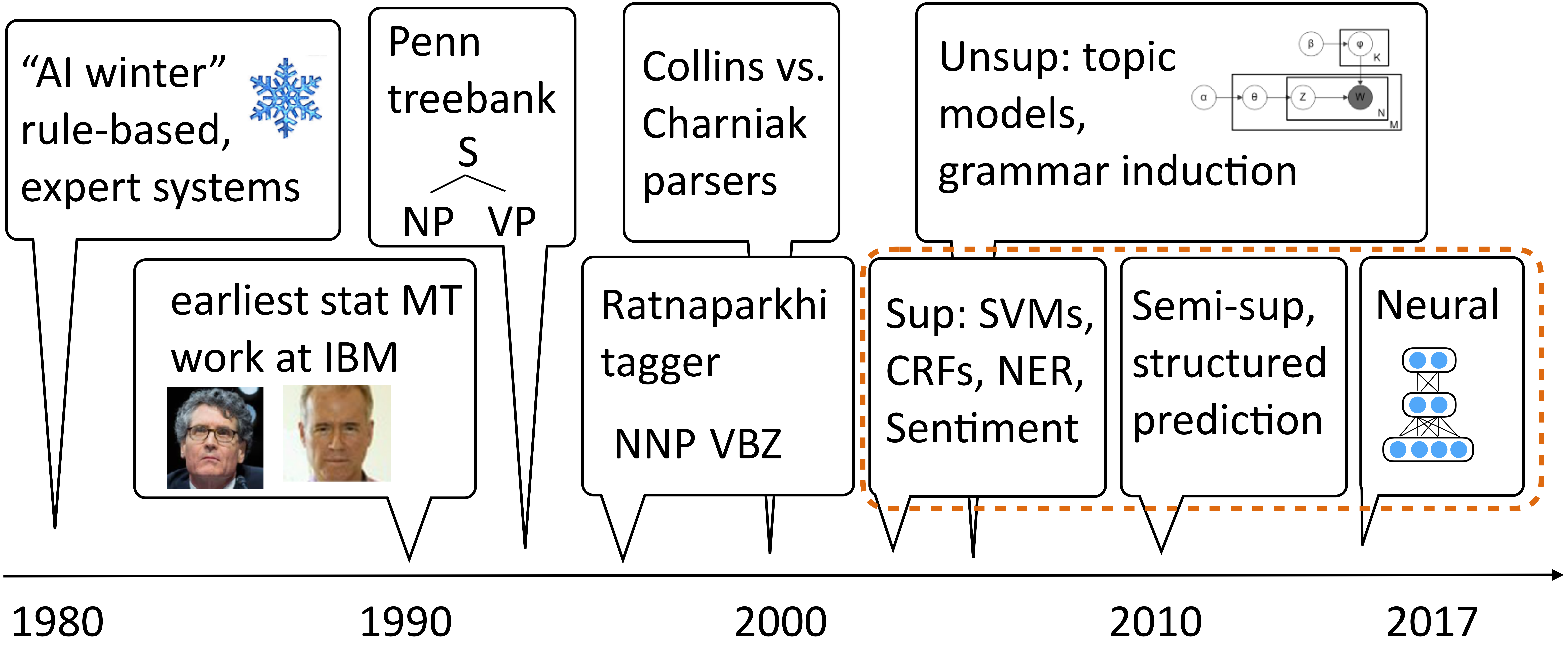
Takeaways

- Why do unsupervised learning when you can annotate a little data and use supervised methods?

- Are these structures even useful anyway?

Structure in NLP

A brief history of (modern) NLP



- What was the role of structured models after all?

Sequential Structure: Analysis

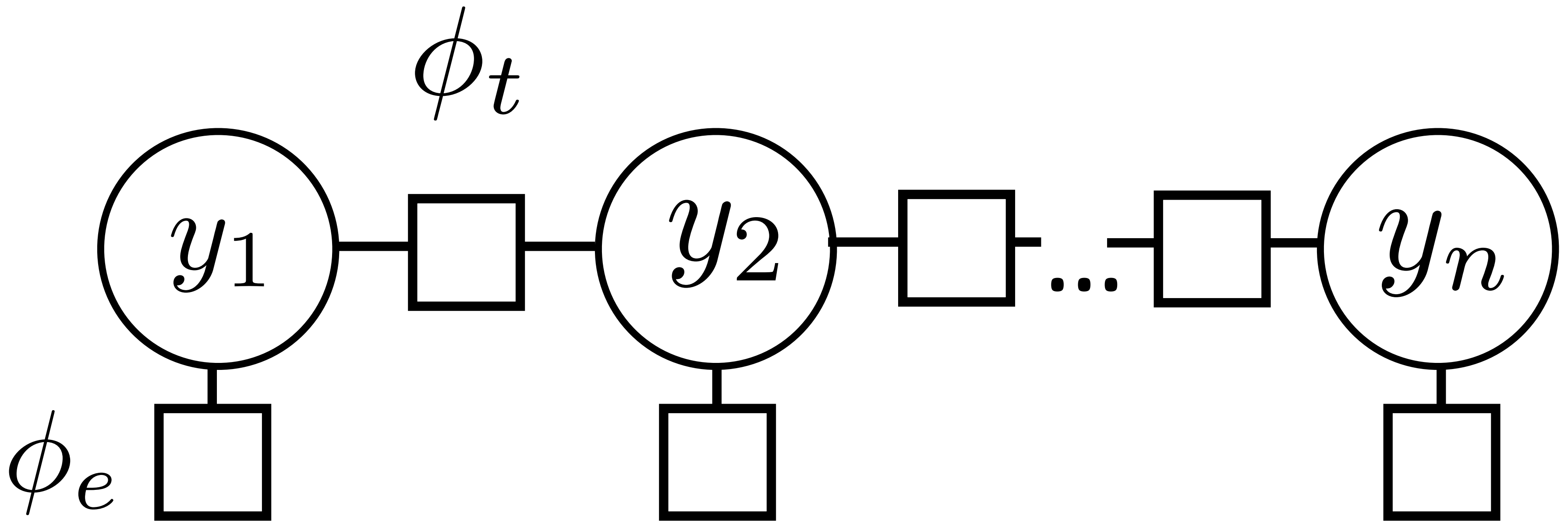
- Language is inherently sequential

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

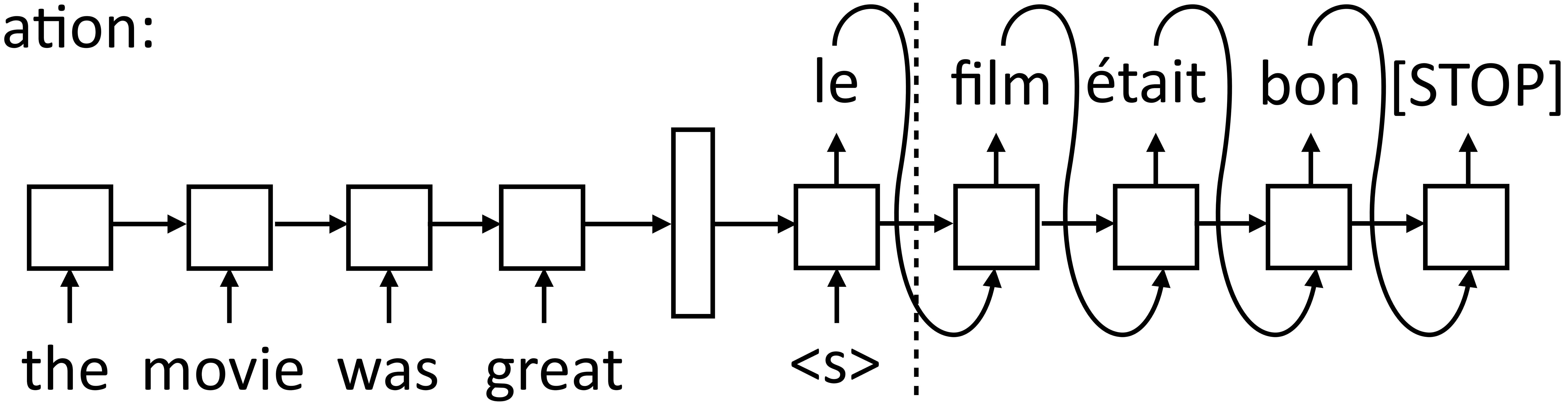
PERSON LOC ORG

- Can do language analysis with sequence models

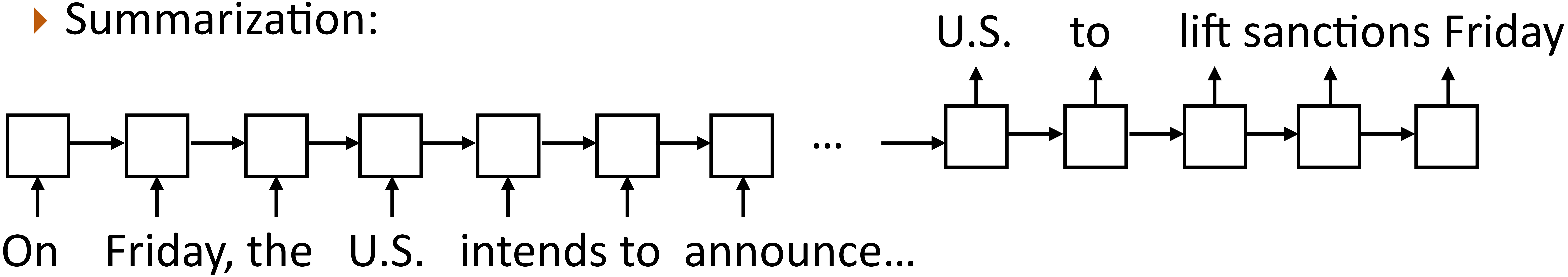


Sequential Structure: Generation

- Translation:



- Summarization:



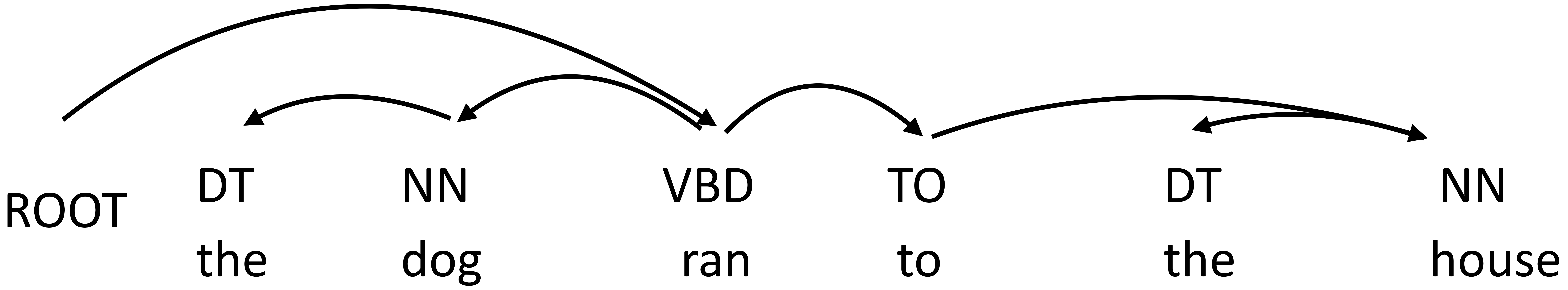
Sequential Structure

- Making a sequence of decisions is the easiest way to predict something structured

- Is the right information always local?

Tree Structure: Analysis

- Parse trees expose sentence structure more directly:

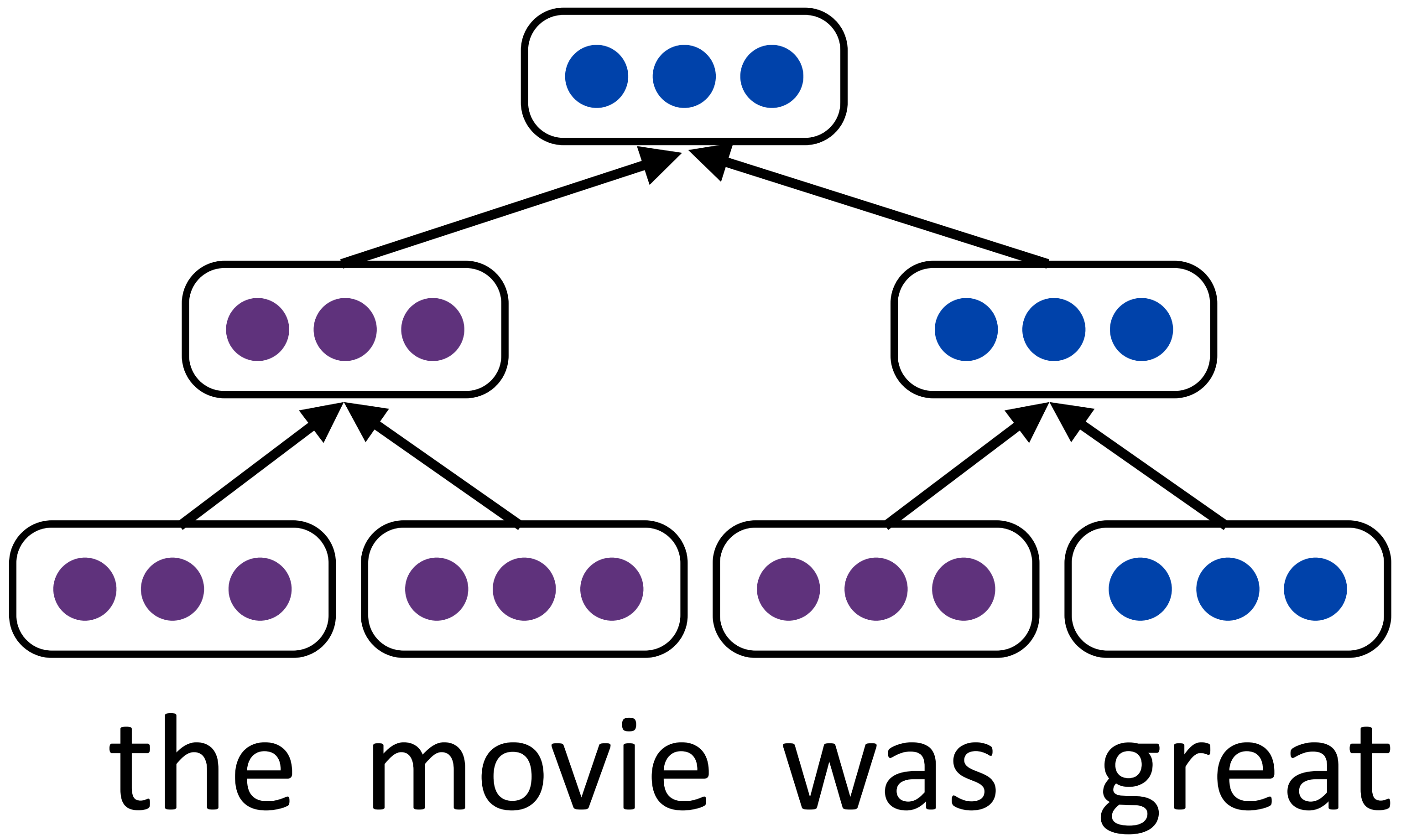


- Semantic roles: (ran, SUBJ=dog, IOBJ=house)

- AMRs that include coreference, etc.

Tree Structure: Analysis

- Useful in combination with neural networks for tasks like sentiment analysis



- How much do explicit trees help?

Tree Structure: Analysis

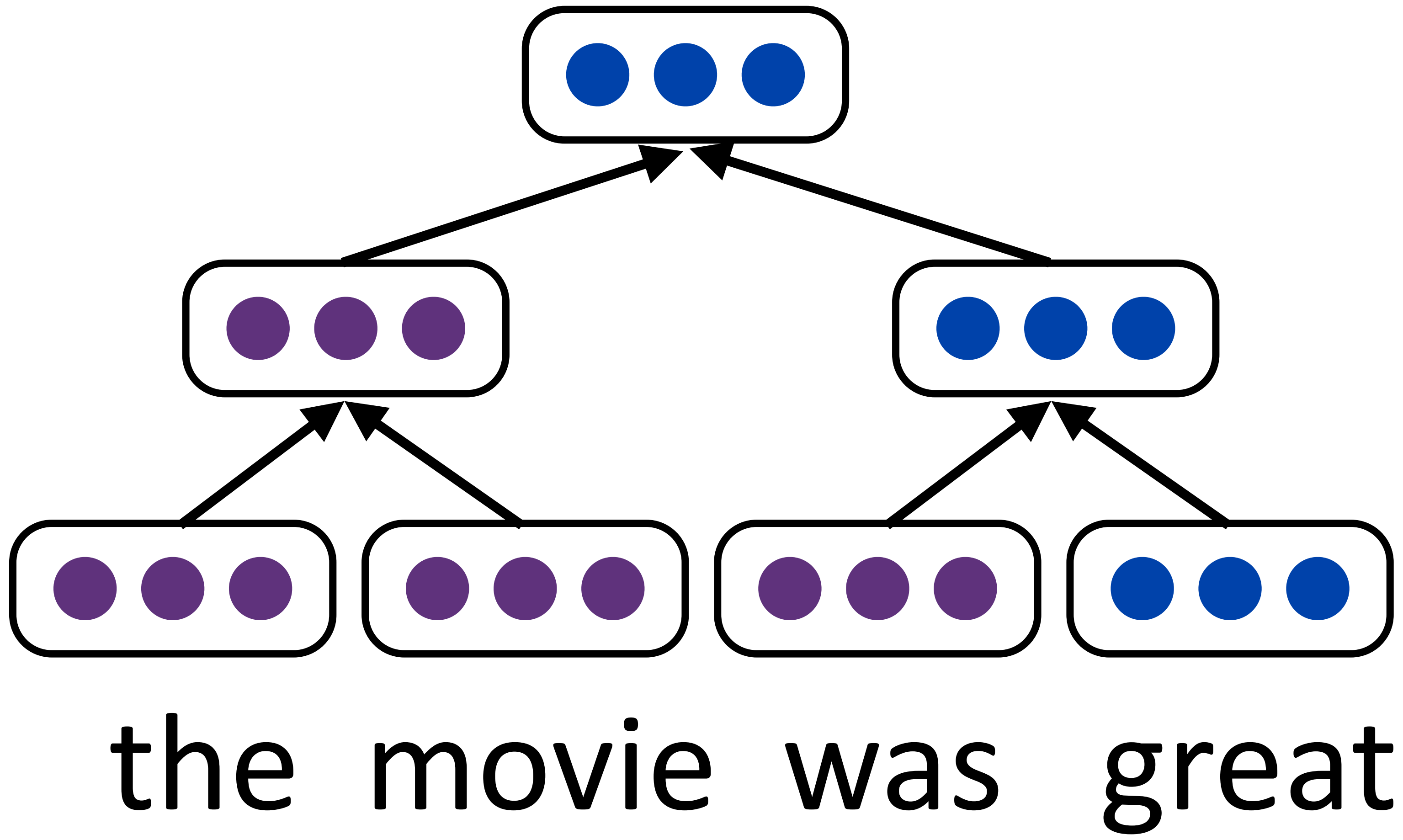
Model	Fine	Binary
RNN (Socher et al. (2011))	43.2	82.4
RNTN (Socher et al. (2013))	45.7	85.4
DRNN (Irsoy & Cardie (2014))	49.8	86.8
RLSTM (Tai et al. (2015))	51.0	88.0
DCNN (Kalchbrenner et al. (2014))	48.5	86.9
CNN-MC (Kim (2014))	47.4	88.1
Bi-LSTM (Tai et al. (2015))	49.1	87.5
LSTMN (Cheng et al. (2016))	47.9	87.0
PVEC (Le & Mikolov (2014))	48.7	87.8
DAN (Iyyer et al. (2014))	48.2	86.8
DMN (Kumar et al. (2016))	52.1	88.6
Kernel NN, $\lambda = 0.5$	51.2	88.6
Kernel NN, gated λ	**53.2**	**89.9**

trees

trees

no trees

no trees

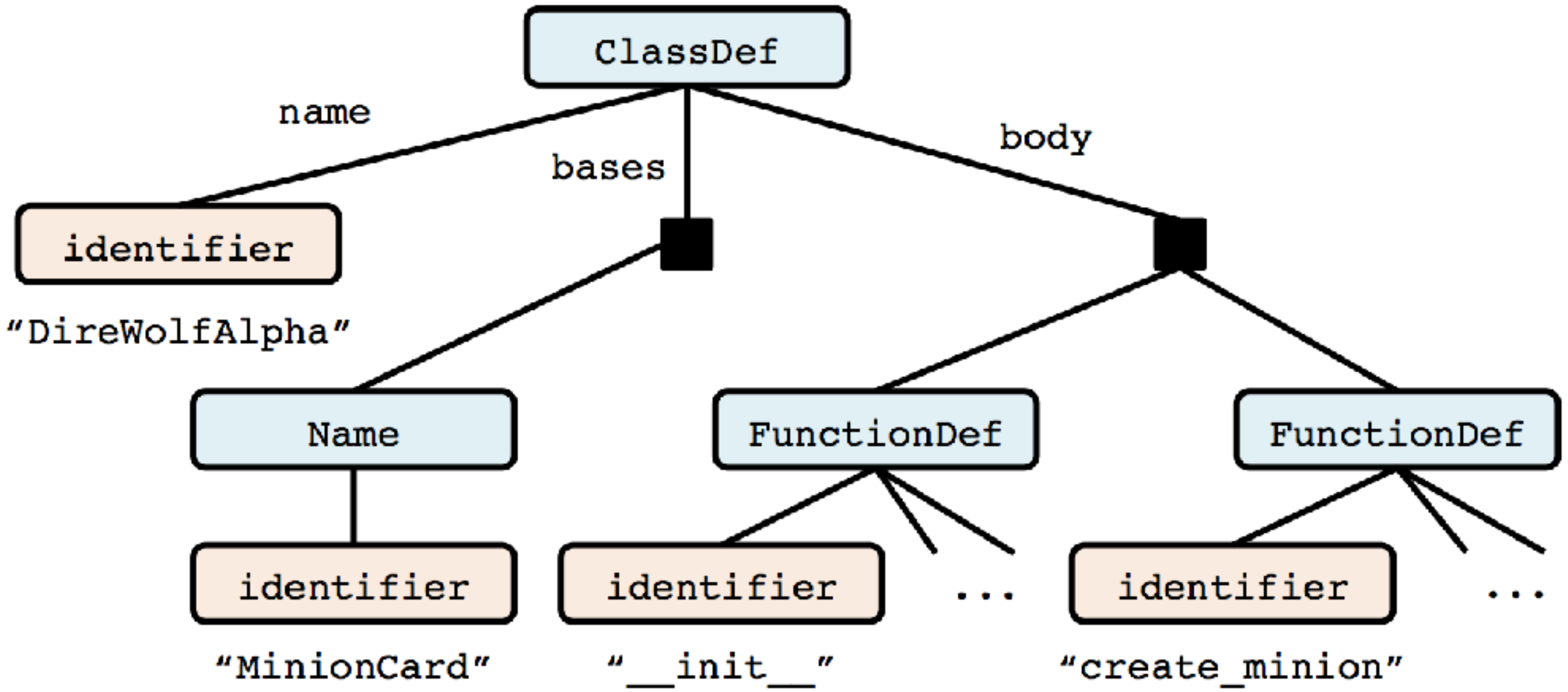


???

Lei et al. (2017)

Tree Structure: Generation

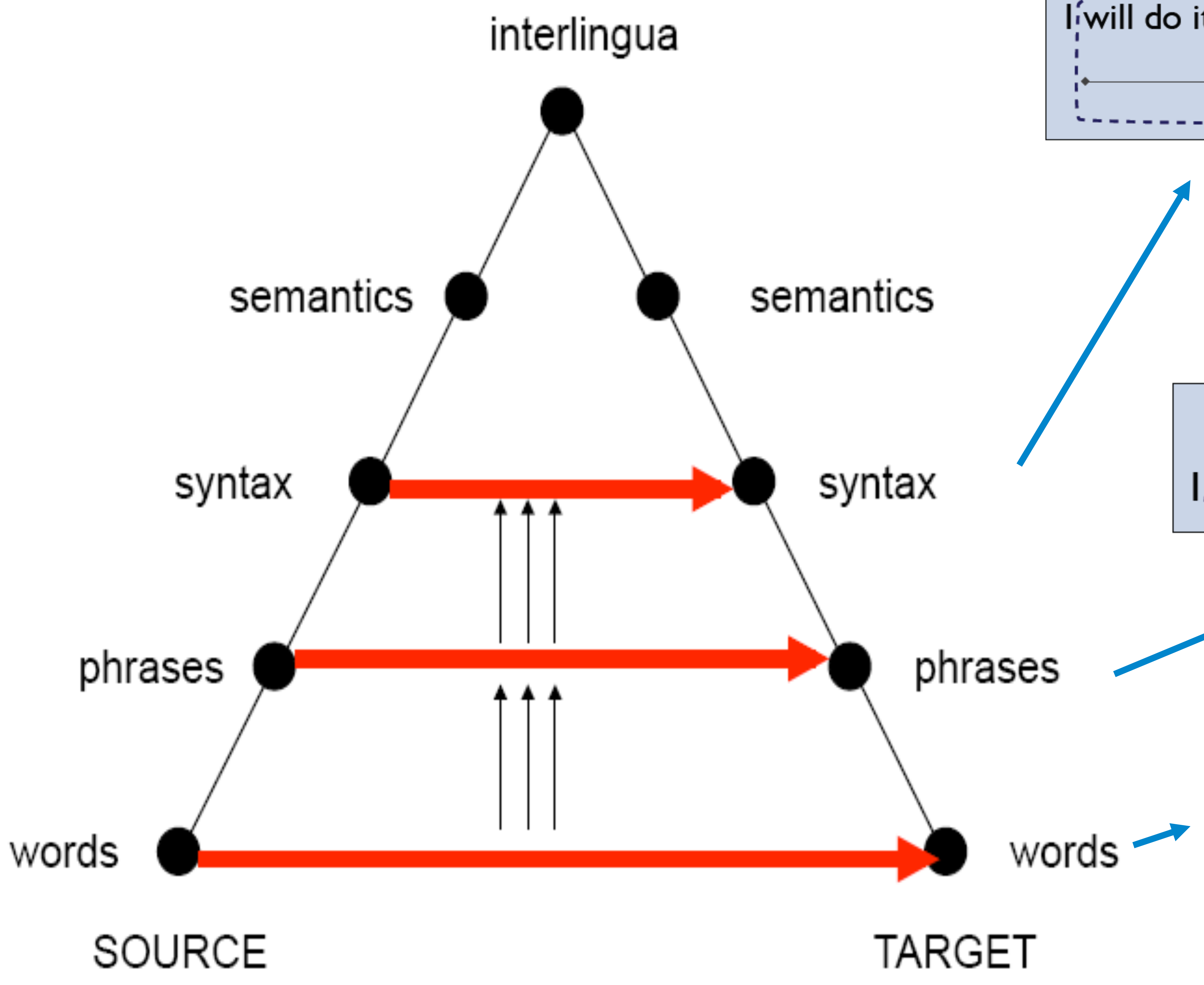
- Generate structured things like source code

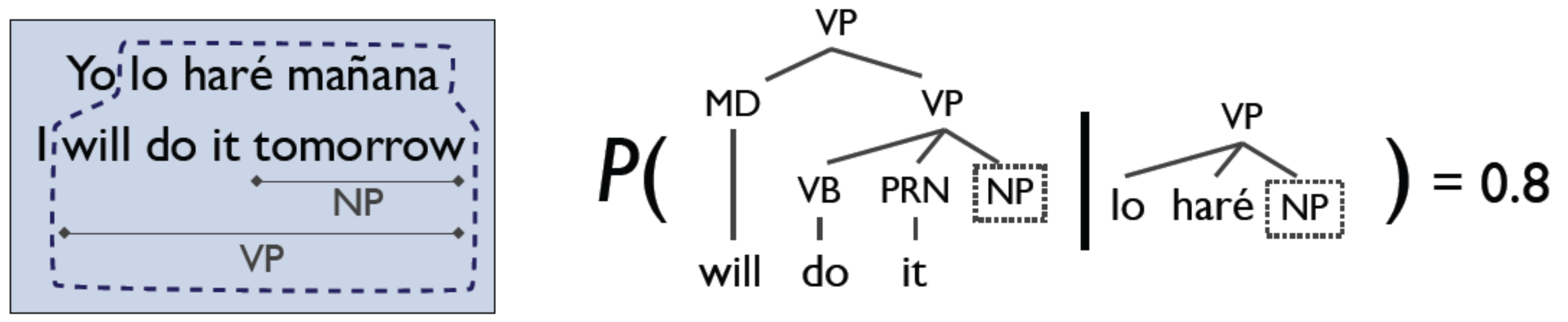


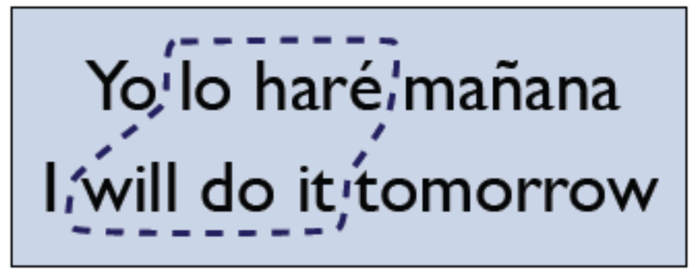
- Generate sentences? Maybe…

Rabinovich et al. (2017)

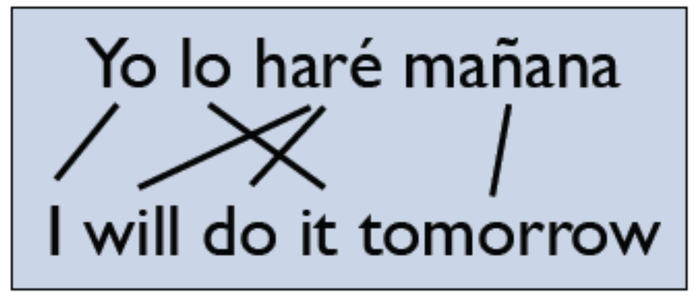
Tree Structure: Generation







English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

- Current best MT systems are arguably word (or even character!) level, but also arguably more abstract…

slide credit: Dan Klein

Higher-level Structure

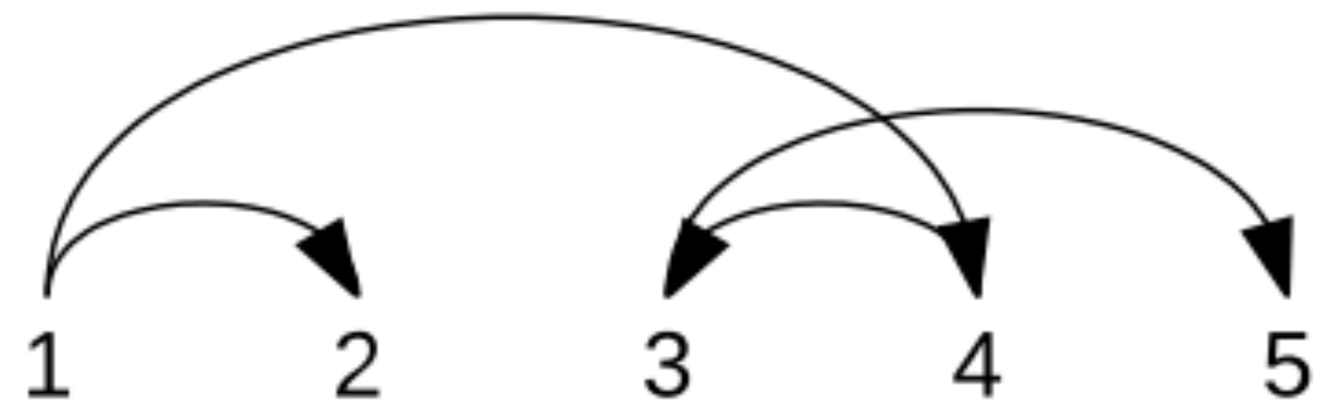
- Main challenge of NLP: how do we achieve bigger-picture understanding?

Higher-level Structure: Documents

- Latent models of discourse structure can help with sentiment analysis

1 great instruction by ryan
2 clean workout facility and friendly people
3 they have a new student membership for 60 per month and classes are mon , weds and fri 6pm 7pm
4 it 's definitely worth money if you want to learn brazilian jiu jitsu
5 i usually go to classes on mondays and fridays , and it 's the best workout i 've had in years

(b)



- Summarization: explicit models of discourse aren't all that useful, but implicit ones might be

Liu and Lapata (2017)

Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences

She's a lovely girl. She has long and black hair. She is quite tall and slim. Her eyes are bright and black. She is 13 years old. She is good at singing. She likes listening to music. She is S.H.E.'s fan . Do you know Conan? He is a little detective .The lovely girl also likes him. Oh, sorry. I forget to tell you who the girl is. It's me. I'm a lovely girl. You can call me Kacely or Kacelin. Now I study at Sunshine Middle School. I'm in Class 1, Grade 7. Every day, I get up at 6:00 a.m. The classes begin at 7 o'clock. I like lunchtime because I can chat with my friends at that time. After school, I usually play badminton with my friends. I like playing badminton and I am good at it. I want to be a superstar when I grow up.

Kacely is a 12-year-old girl. She currently goes to Sunshine Middle School .

Q: Kacely is a ____?

A) student
B) teacher
C) principal
D) parent

She → Kacely — coreference

Kacely goes to school — parsing

Kacely goes to school ENTAILS Kacely is a **student** — entailment

Higher-level Structure: Dialogue

Find me a good sushi restaurant in Chelsea

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

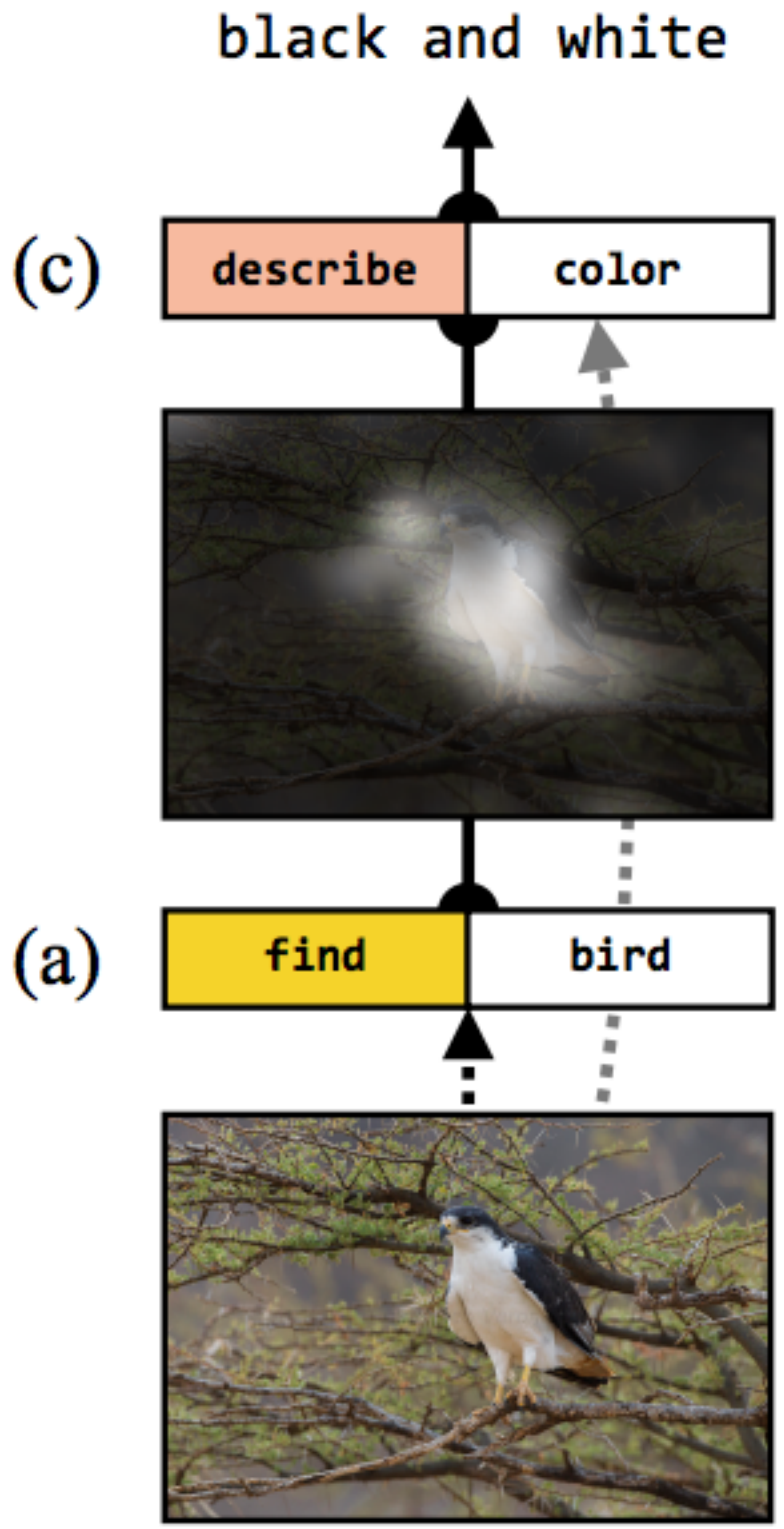
- What do we need to track to answer questions like this? How do we track it?

Higher-level Structure: Grounding

- Language is learned through interacting with an environment

- "Bird", "white" — these concepts are anchored in visual features, can compose to get "white bird"

- Great for physically anchored things — how do we learn a representation of "democracy"?



what color is the bird?

Andreas et al. (2016)

How do we get here?

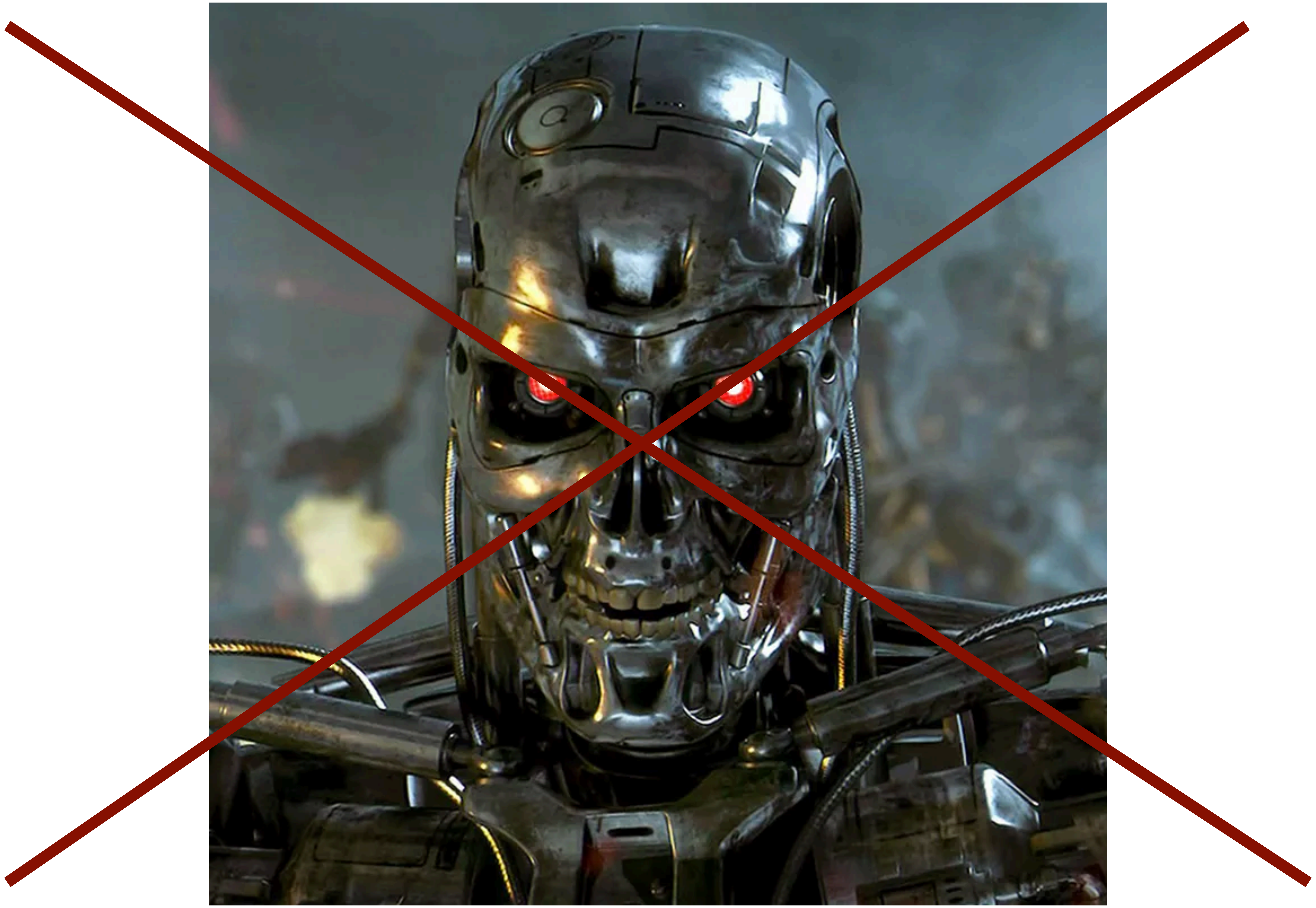
- Neural networks let us learn from data in an end-to-end way, very powerful learners

- Structure imposes inductive biases in these networks

- Leverage model structure to do reasoning (discrete reasoning?)

- Need to solve all of these challenges: ground language in the world and leverage information across whole dialogues/documents — otherwise systems are inherently limited

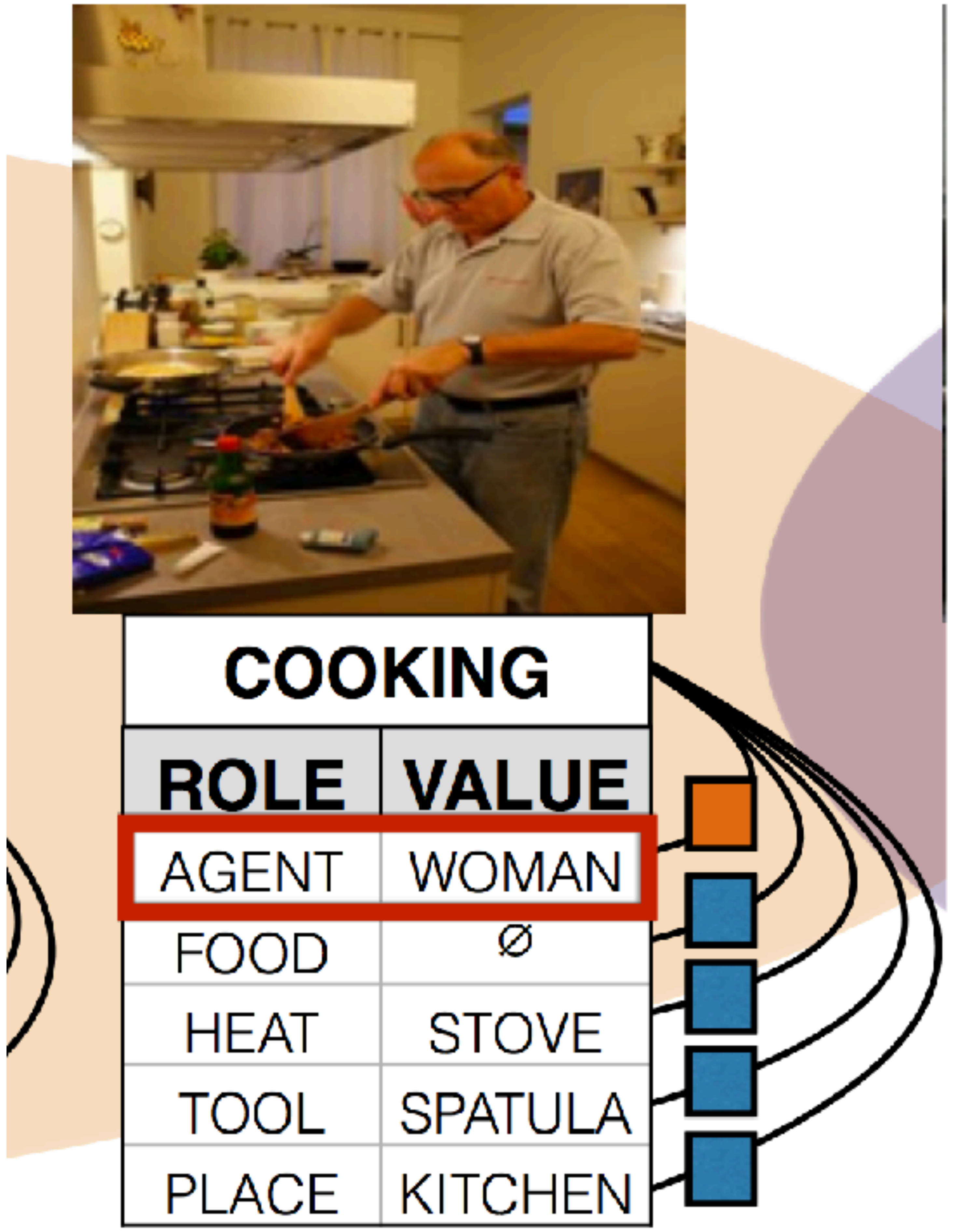
Ethics in NLP — what can go wrong?



What can actually go wrong?

Bias Amplification

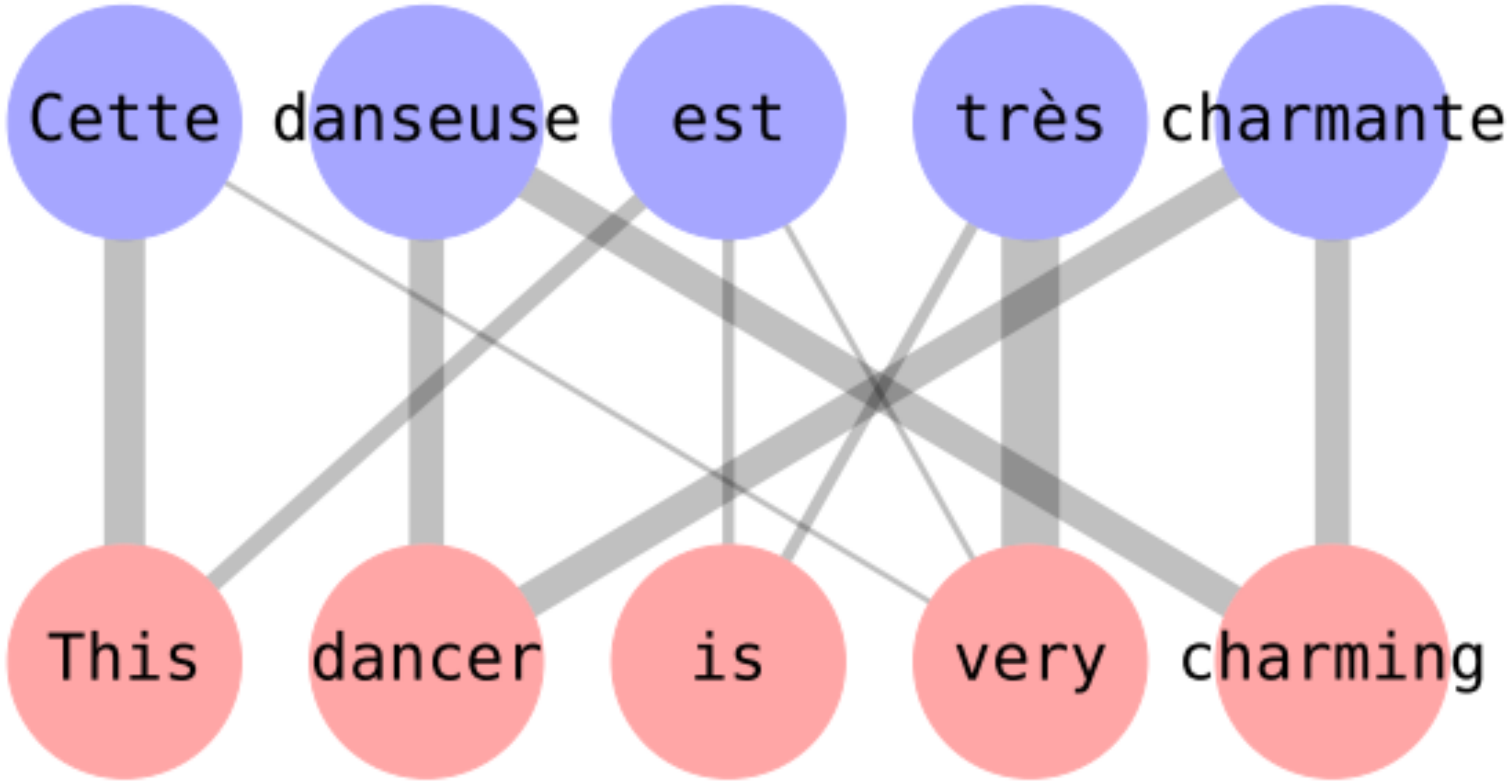
- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias
- Can we constrain models to avoid this while achieving the same predictive accuracy?
- Place constraints on proportion of predictions that are men vs. women?



Zhao et al. (2017)

Bias Amplification

- Harder to quantify this for machine translation
- “dancer” is assumed to be female in the context of the word “charming”… but maybe that reflects how language is used?



Alvarez-Melis and Jaakkola (2017)

Exclusion

- Most of our annotated data is English data, especially newswire

- What about:

    Dialects?

    Other languages? (Non-European/CJK)

    Codeswitching?

Unethical Use

- Surveillance applications?
- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	—
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

- What if these were undetectable?

Dangers of Automatic Systems

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD TECH POLITICS

Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'

14

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

Slide credit: The Verge

Dangers of Automatic Systems

Translations of **gay**

adjective

homosexual	homosexual, gay, camp
alegre	cheerful, glad, joyful, happy, merry, gay
brillante	bright, brilliant, shiny, shining, glowing, glistening
vivo	live, alive, living, vivid, bright, lively
vistoso	colorful, ornate, flamboyant, colourful, gorgeous
jovial	jovial, cheerful, cheery, gay, friendly
gayo	merry, gay, showy

noun

el homosexual	$\boxed{\text{homosexual, gay, poof, queen, faggot, fagot}}$
el jovial	gay

‣ Offensive terms

Slide credit: allout.org

Dangers of Automatic Systems

"Instead of relying on algorithms, which we can be accused of manipulating for our benefit, we have turned to machine learning, an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie."

\- Maciej Cegłowski

Slide credit: Sam Bowman