THE OHIO STATE UNIVERSITY

CSE 5525: Foundations of Speech and Language Processing

Wrapup and Ethics

Huan Sun (CSE@OSU)

Slides were largely adapted from Prof. Greg Durrett @ UT Austin.

Administrivia

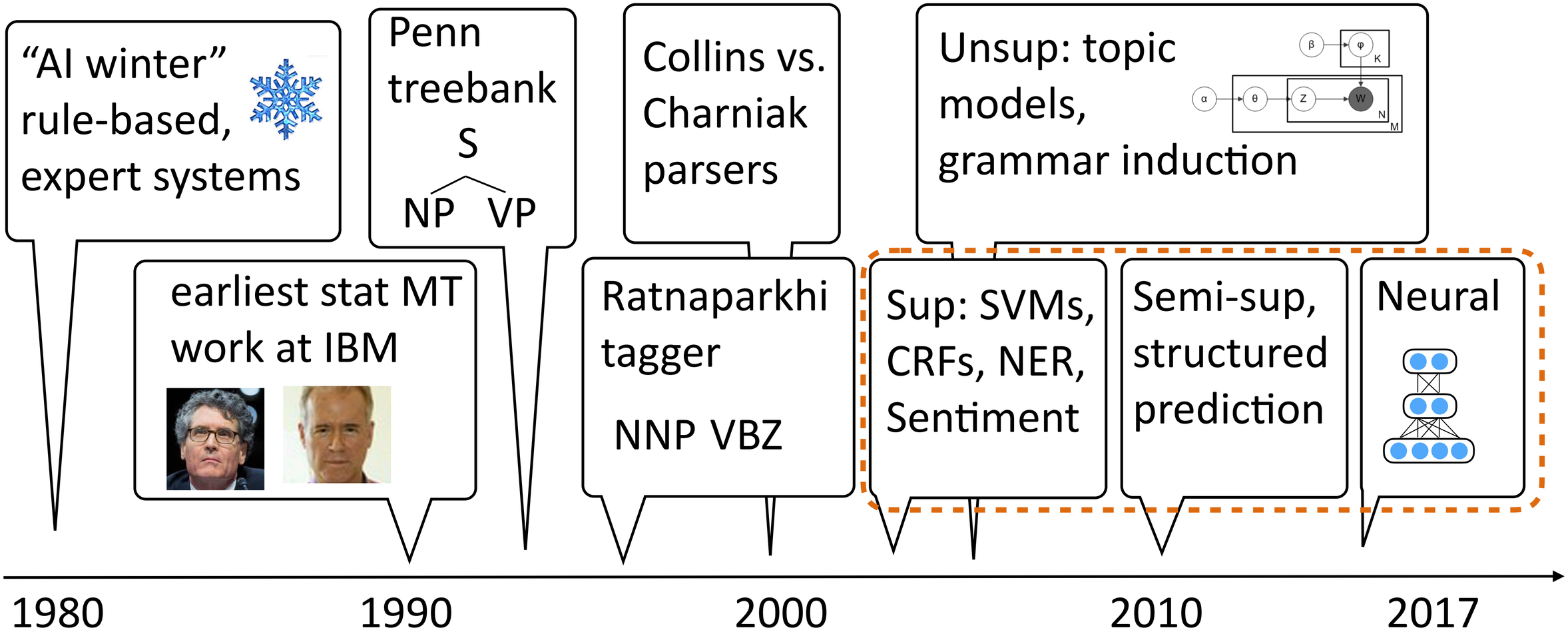
- SEI surveys
- HW3 graded (grades for grads will be uploaded later)
- Final project presentations on Dec 2 and 4.
  - See Carmen announcement to sign up
  - 10-minute presentation (including QA)
  - Can be “work in progress”, but should at least have preliminary results
  - Final reports due on December 6; no slip days
    - The format of your final report, e.g., https://arxiv.org/pdf/2010.12800.pdf

This Lecture

- Wrapup and current challenges

- Ethics in NLP/ML

A brief history of (modern) NLP



- What different model structures did we consider?

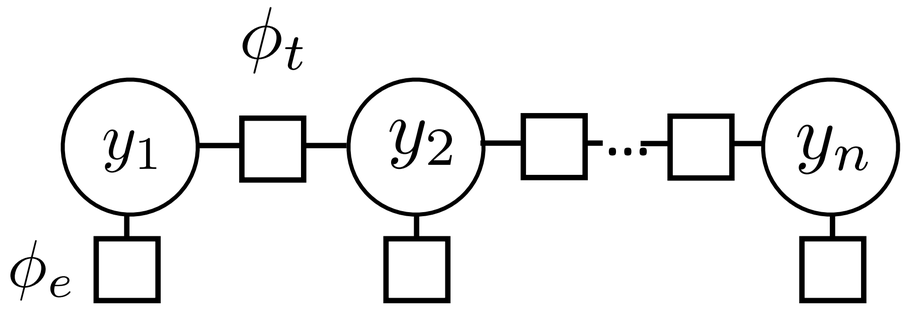
Sequential Structure: Analysis

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 .

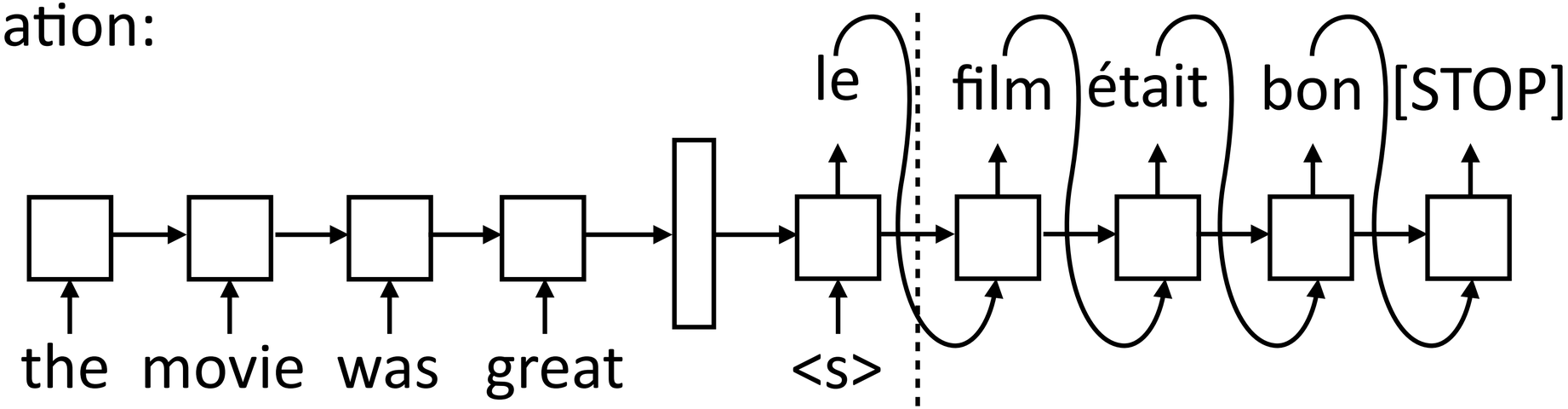
PERSON LOC ORG

- Can do language analysis with sequence models

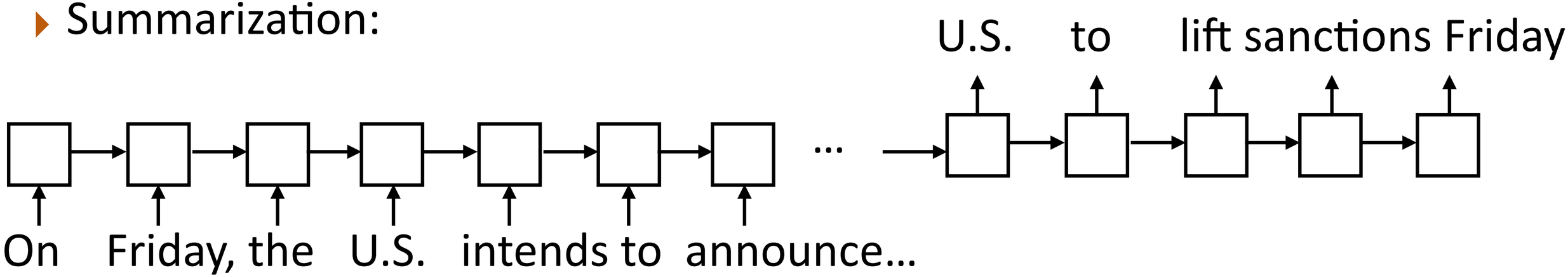


Sequential Structure: Generation

- Translation:

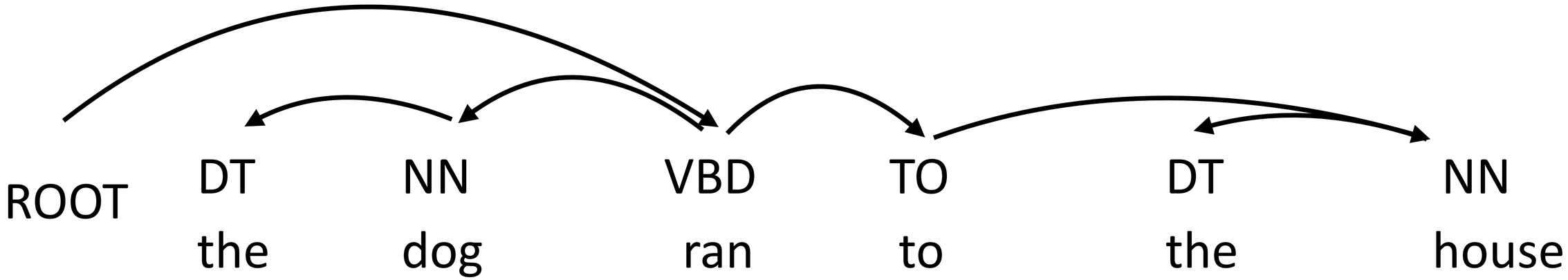


- Summarization:



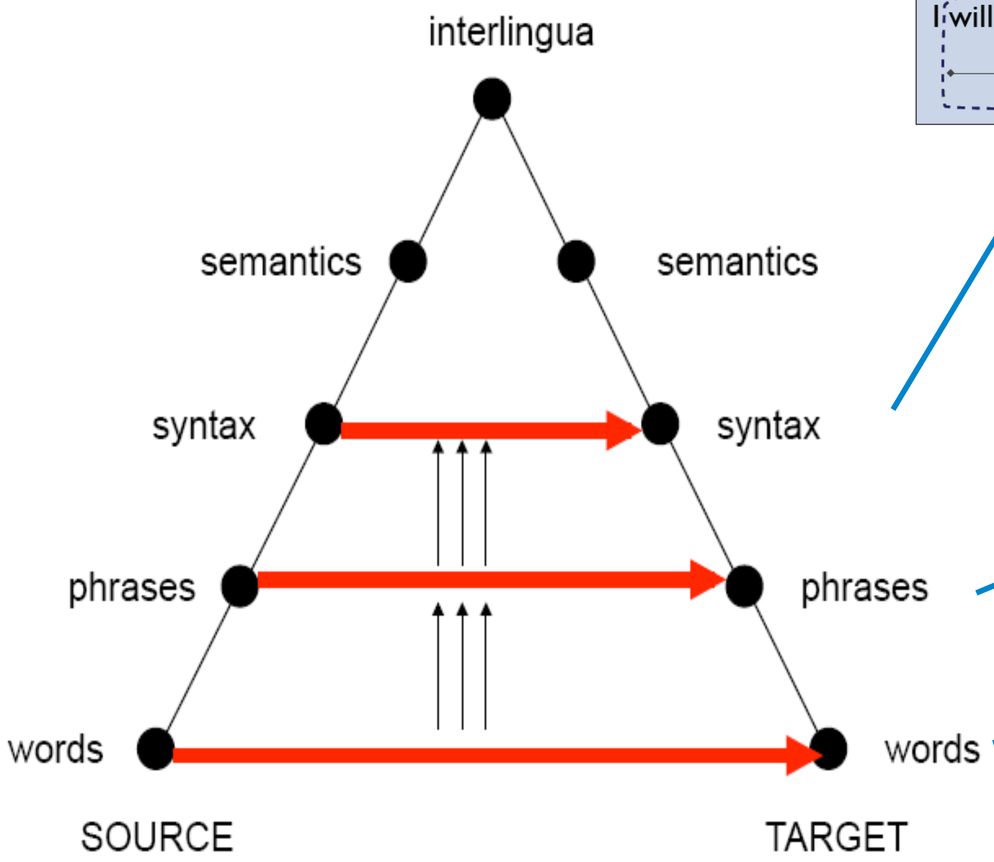
Tree Structure: Analysis

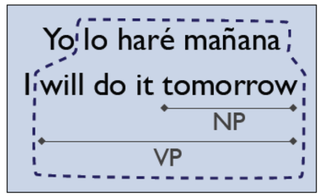
- Parse trees expose and localize the right information more directly:

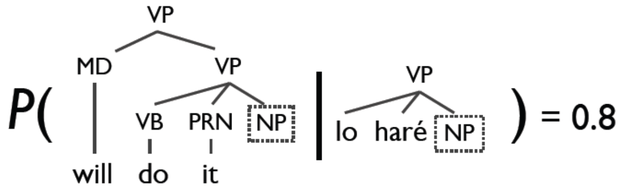


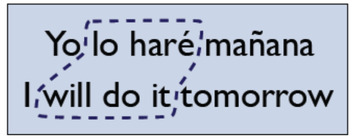
- Semantic roles: (ran, SUBJ=dog, IOBJ=house)

Tree Structure: Generation

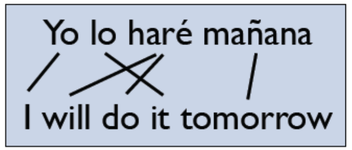








English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



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

slide credit: Dan Klein

What can we do (well)?

- QA, summarization, machine translation, …
  - …for domains where we have 10k+ or 100k+ examples (10M+ for MT)
  - …and the input/output correspondence isn't too complicated
- Neural networks let us learn from data in an end-to-end way, very powerful learners…but there are limits to what they can learn

What can’t we do (well)?

- Generalize models to new domains

Q: Is Hirschsprung disease a Mendelian or a multifactorial disorder?

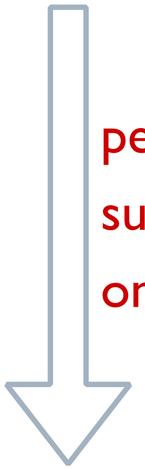
> Coding sequence mutations in RET, GDNF, EDNRB, EDN3, and SOX10 are involved in the development of Hirschsprung disease. The majority of these genes was shown to be related to Mendelian syndromic forms of Hirschsprung's disease, whereas the non-Mendelian inheritance of sporadic non-syndromic Hirschsprung disease proved to be complex; involvement of multiple loci was demonstrated in a multiplicative model.
>
> The non-Mendelian inheritance of sporadic non-syndromic Hirschsprung's disease proved to be complex; involvement of multiple loci was demonstrated in a multiplicative model

- Arguably humans can’t always do this either, but we can be taught quickly! How could a machine learn from a textbook?
- BERT can help, but there’s a long way to go

Example: Clinical Reading Comprehension

Model	emrQA Relation	
	EM	F1
DocReader	86.94	94.85
ClinicalBERT	85.33	93.06

Train on emrQA
test on emrQA



performance substantially drops on a different corpus

Overall	
EM	F1
40.00	53.27
38.00	60.19

Train on emrQA
test on MIMIC-III

[Yue et al., ACL'20]

Example: Cross-database Semantic Parsing

Dataset	Metric	# Examples	Our best	–WikiSQL	–SQL^{UF}	– Value copying	Empty Prior
ATIS	*Execution*	289 (486)	0.8 (11.9)	0.5 (11.9)	0.8 (11.9)	0.1 (10.8)	0.0 (11.9)
GeoQuery		532 (598)	41.6 (40.0)	35.6 (35.0)	34.7 (33.4)	2.2 (5.6)	0.0 (4.0)
Restaurants		27 (378)	3.7 (45.2)	3.7 (46.3)	0.0 (46.6)	0.0 (51.1)	0.0 (51.6)
Academic		180 (196)	8.2 (12.1)	6.1 (9.4)	5.7 (9.0)	2.8 (7.7)	0.0 (4.1)
IMDB		107 (131)	24.6 (33.3)	24.3 (32.3)	23.1 (32.3)	0.0 (14.3)	0.0 (13.0)
Yelp		54 (128)	19.8 (49.2)	16.7 (47.9)	14.8 (47.9)	4.9 (53.1)	0.0 (41.4)
Scholar		394 (599)	0.5 (6.8)	0.4 (7.4)	0.5 (8.6)	0.2 (7.8)	0.0 (9.3)
Advising		309 (2858)	2.3 (35.2)	1.2 (35.7)	1.4 (37.3)	0.0 (38.0)	0.0 (38.3)
Spider	*Execution*	1034	69.0	68.4	65.1	33.9	4.7
	Exact Set Match		65.0	65.1	60.5	54.1	–

Train on Spider; test on 8 other datasets

[Suhr et al., ACL'20]

Example: Cross-database Semantic Parsing

	Models	Spider-Realistic	ATIS	GeoQuery	Restaurants	Academic	IMDB	Yelp	Scholar	Advising
	# Examples	508	289	532	27	180	107	54	394	309
Schema Only	Suhr et al. (2020)	-	0.8 (0.5)	41.6 (35.6)	3.7 (3.7)	8.2 (6.1)	24.6 (24.3)	**19.8 (16.7)**	0.5 (0.4)	2.3 (1.2)
	RAT-SQL *w/o* value linking									
	w. $\text{BERT}_{\text{LARGE}}$	52.4 ± 0.7 (46.9)	2.1 ± 0.6	41.2 ± 11.6	0.0 ± 0.0	5.9 ± 2.1	26.5 ± 5.0	12.3 ± 1.7	0.8 ± 0.4	1.6 ± 0.4
	w. StruG (Human Assisted)	57.8 ± 0.6 (53.3)	2.2 ± 0.2	45.5 ± 1.8	11.1 ± 9.1	**14.8 ± 5.0**	**37.1 ± 1.8**	15.4 ± 0.9	4.3 ± 1.7	**2.2 ± 0.4**
	w. StruG (Automatic)	**60.3 ± 0.7 (54.9)**	**2.2 ± 0.2**	**50.9 ± 4.0**	**40.7 ± 5.2**	12.4 ± 1.9	35.5 ± 2.0	13.0 ± 2.6	**5.4 ± 0.7**	1.0 ± 0.3
Content Used	RAT-SQL									
	w. $\text{BERT}_{\text{LARGE}}$	62.1 ± 1.3 (58.1)	2.3 ± 0.2	47.3 ± 3.7	37.0 ± 18.9	15.6 ± 2.0	21.8 ± 1.6	16.0 ± 3.1	3.4 ± 1.4	6.4 ± 2.3
	w. GRAPPA	- (59.3)								
	w. StruG (Human Assisted)	**65.7 ± 0.7 (62.2)**	**5.5 ± 1.1**	**59.5 ± 3.2**	40.7 ± 13.9	18.7 ± 2.1	26.8 ± 2.9	**21.6 ± 2.3**	**6.3 ± 1.8**	**6.9 ± 0.6**
	w. StruG (Automatic)	65.3 ± 0.7 (62.2)	2.8 ± 0.7	57.5 ± 0.2	**44.4 ± 32.7**	**20.2 ± 1.6**	**30.2 ± 5.8**	18.5 ± 1.5	6.1 ± 0.5	5.2 ± 0.5

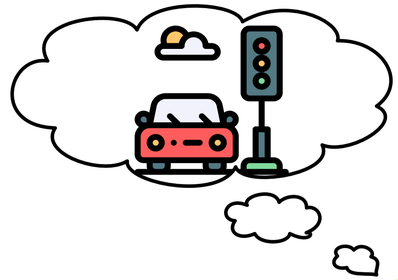
Train on Spider; test on other datasets

[Structure-Grounded Pretraining for Text-to-SQL, arXiv'20]

What can’t we do?

- Commonsense reasoning

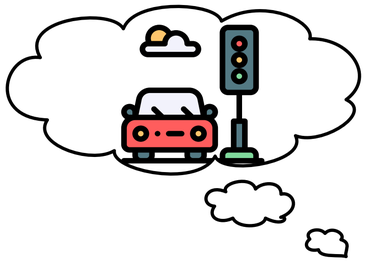
Definition of Common Sense



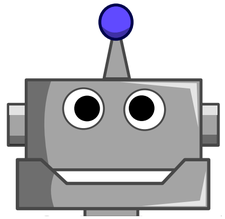
- the basic level of **practical knowledge** and **reasoning**
- concerning **everyday situations** and **events**
- that are **commonly** shared among **most** people.

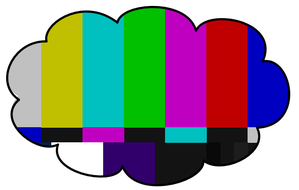
For example, it's ok to keep the closet door open,
but it's not ok to keep the fridge door open,
as the food inside might go bad.

Slide credit: Yejin Choi



Essential for humans to live and interact with each other in a reasonable and safe way.



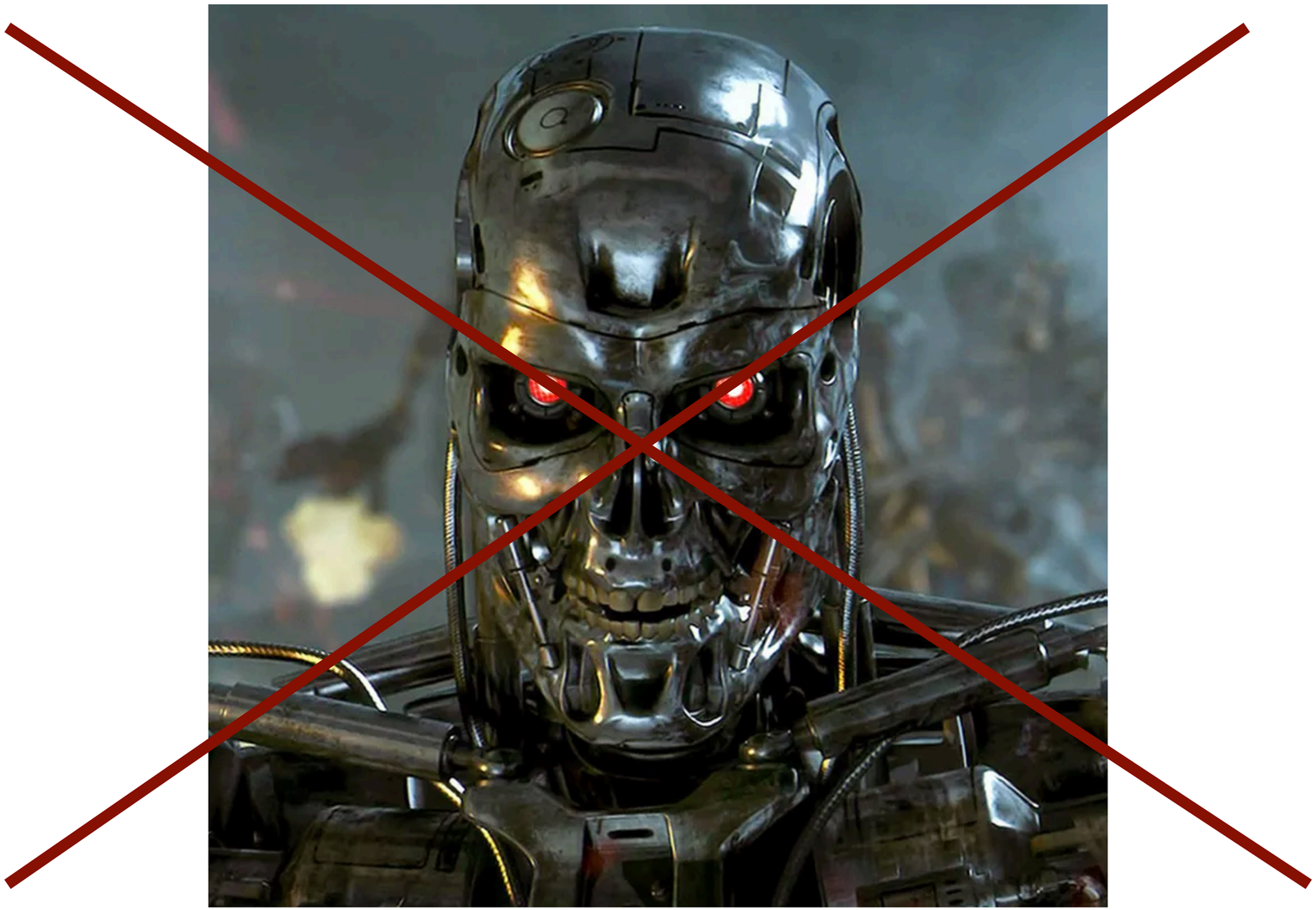


Essential for AI to understand human needs and actions better

For example, it's ok to keep the closet door open,
but it's not ok to keep the fridge door open,
as the food inside might go bad.

Slide credit: Yejin Choi

Ethics in NLP/AI



What can actually go wrong?

Machine-learned NLP Systems

- Aggregate textual information to make predictions
- Hard to know why some predictions are made
- More and more widely use in various applications/sectors

Machine-learned NLP Systems

- Aggregate textual information to make predictions
- Hard to know why some predictions are made
- More and more widely use in various applications/sectors

- What are the risks here?
  - …of certain applications?
    - IE / QA / summarization?
    - MT?
    - Dialog?

Machine-learned NLP Systems

- Aggregate textual information to make predictions
- Hard to know why some predictions are made
- More and more widely use in various applications/sectors

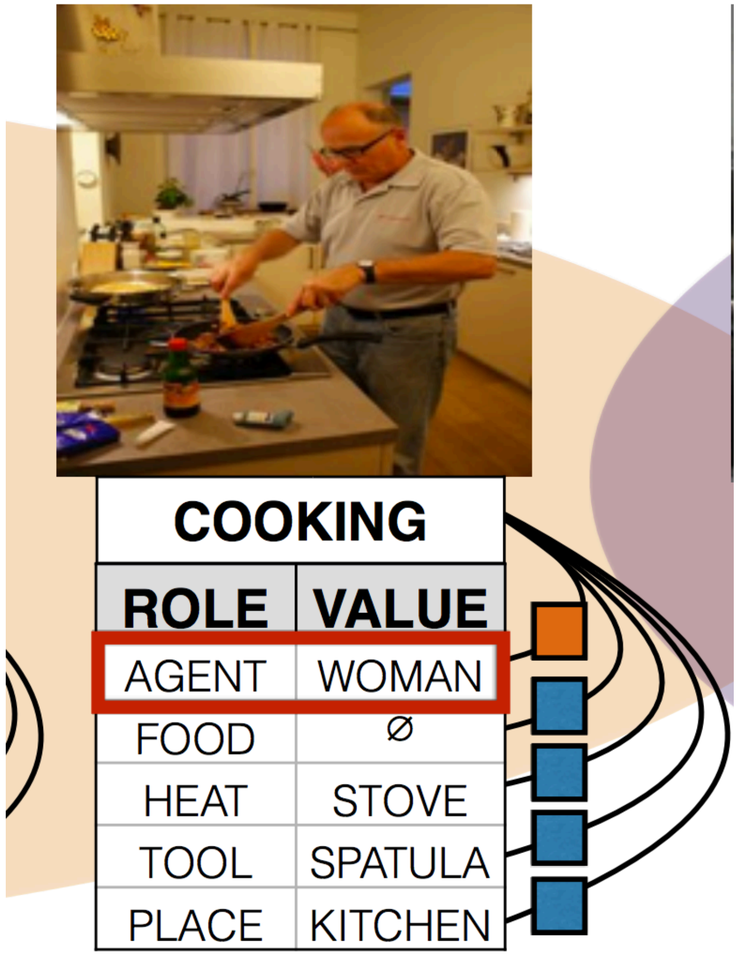
- What are the risks here?
  - …of certain applications?
    - IE / QA / summarization?
    - MT?
    - Dialog?
  - …of machine-learned systems?
  - …of deep learning specifically?

Broad Areas

- Bias amplification: systems exacerbate real-world bias rather than correct for it
- Exclusion: underprivileged users are left behind by systems
- Dangers of automatic systems: automating things in ways we don’t understand is dangerous
- Unethical use: powerful systems can be used for bad ends

Bias Amplification

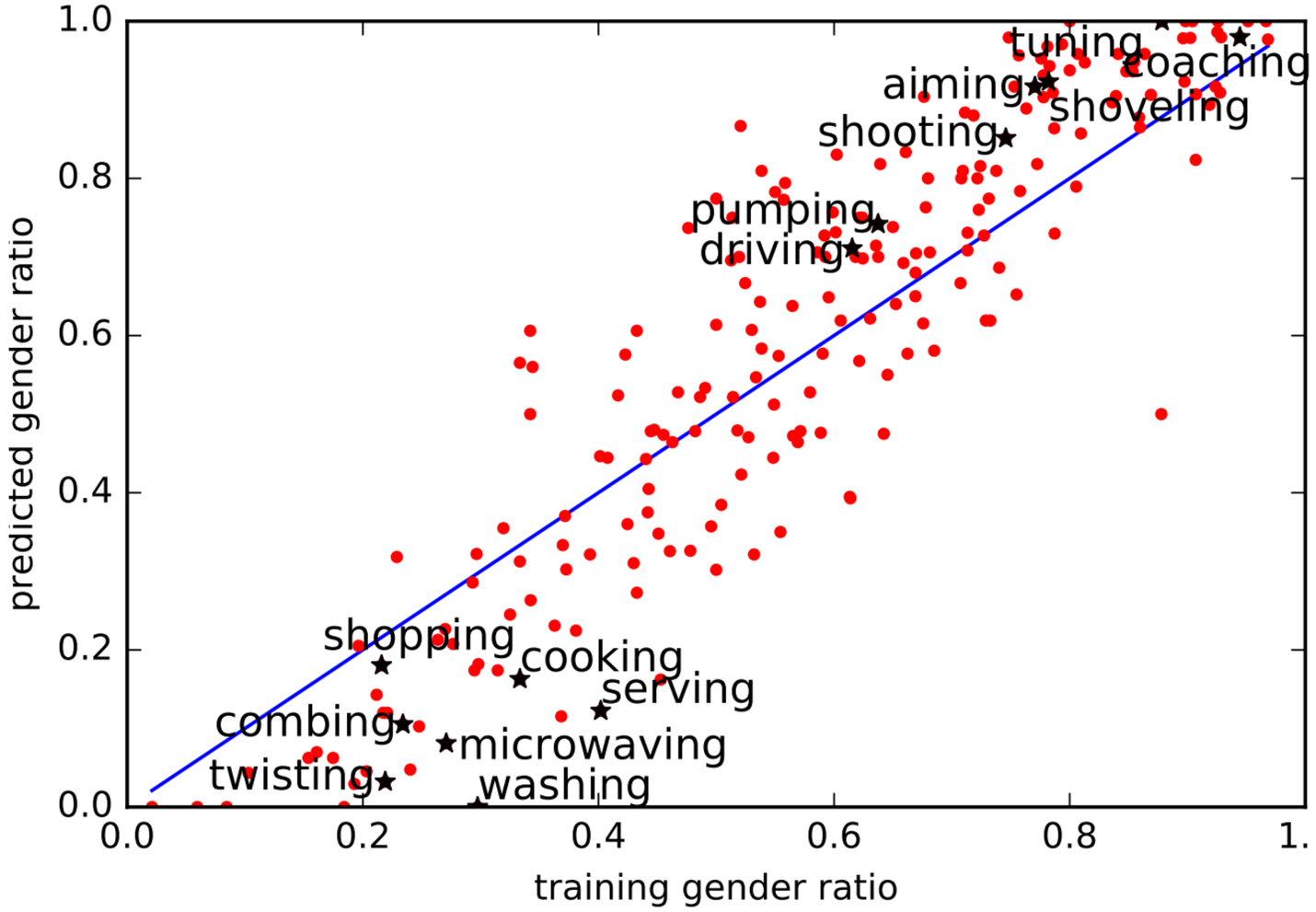
- Bias in data: 67% of cooking images in training have woman in the agent role; but model predicts woman for 84% of cooking images at test time — amplifies bias



Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of cooking images in training have woman in the agent role; but model predicts woman for 84% of cooking images at test time — amplifies bias



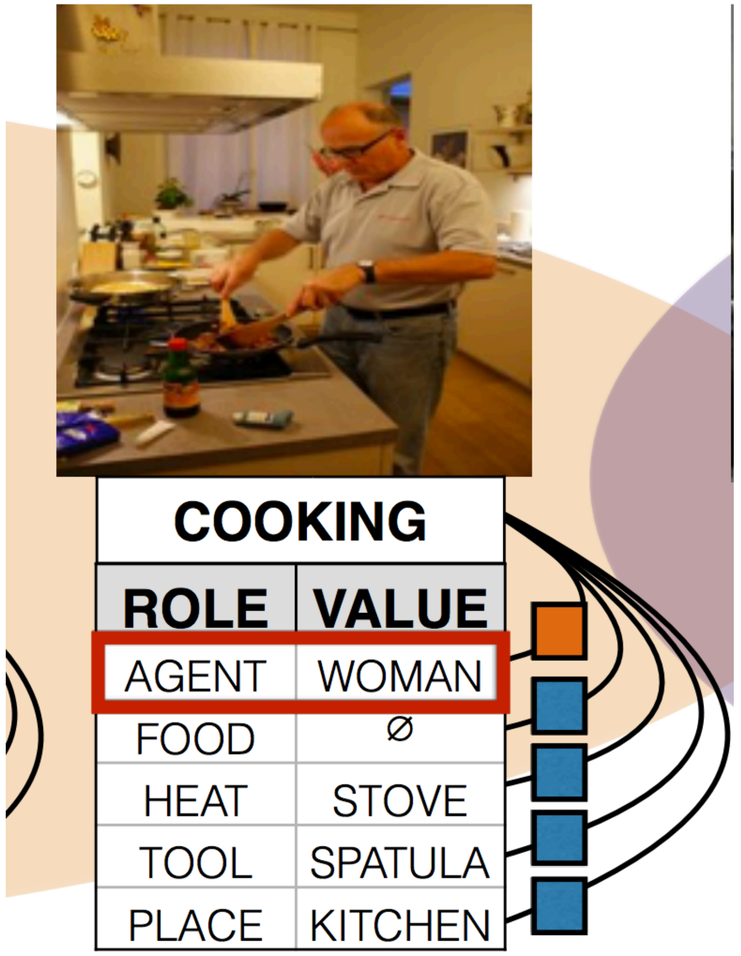
Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of cooking images in training have woman in the agent role; but model predicts woman for 84% of cooking images 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

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

Maximize score of predictions…

f(y, i) = score of predicting y on ith example

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

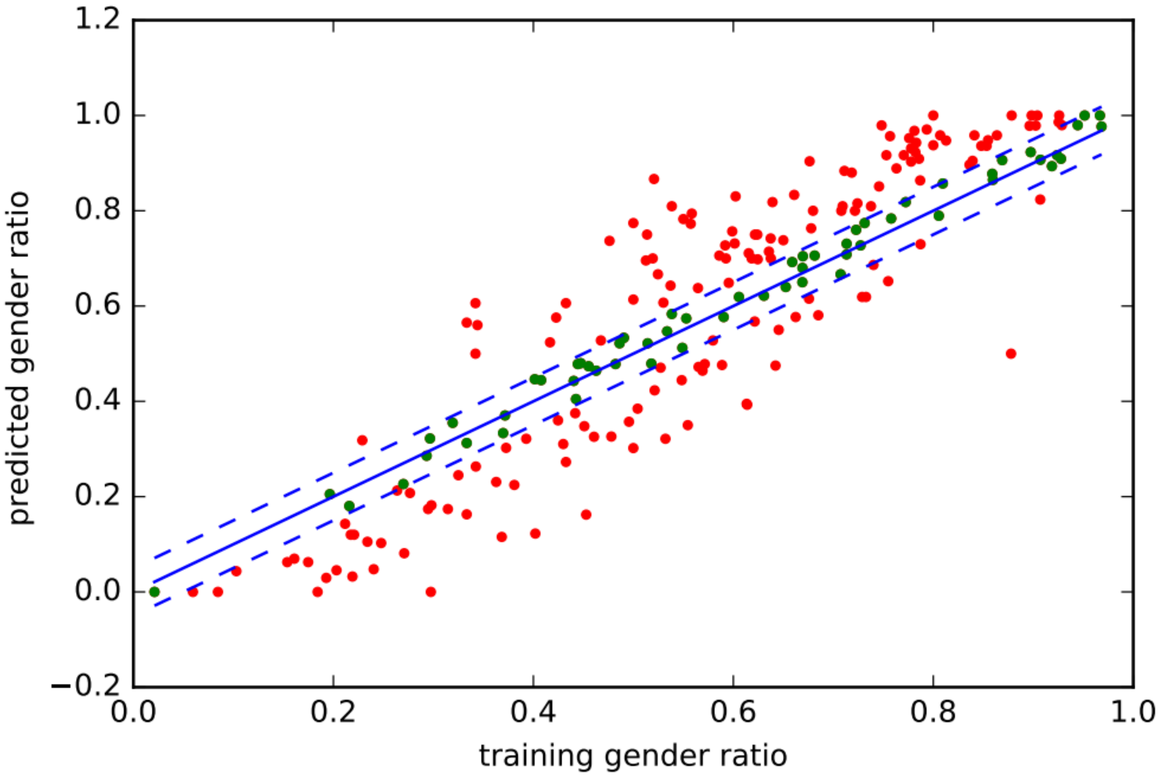
…subject to bias constraint

- Constraints: male prediction ratio on the test set has to be close to the ratio on the training set

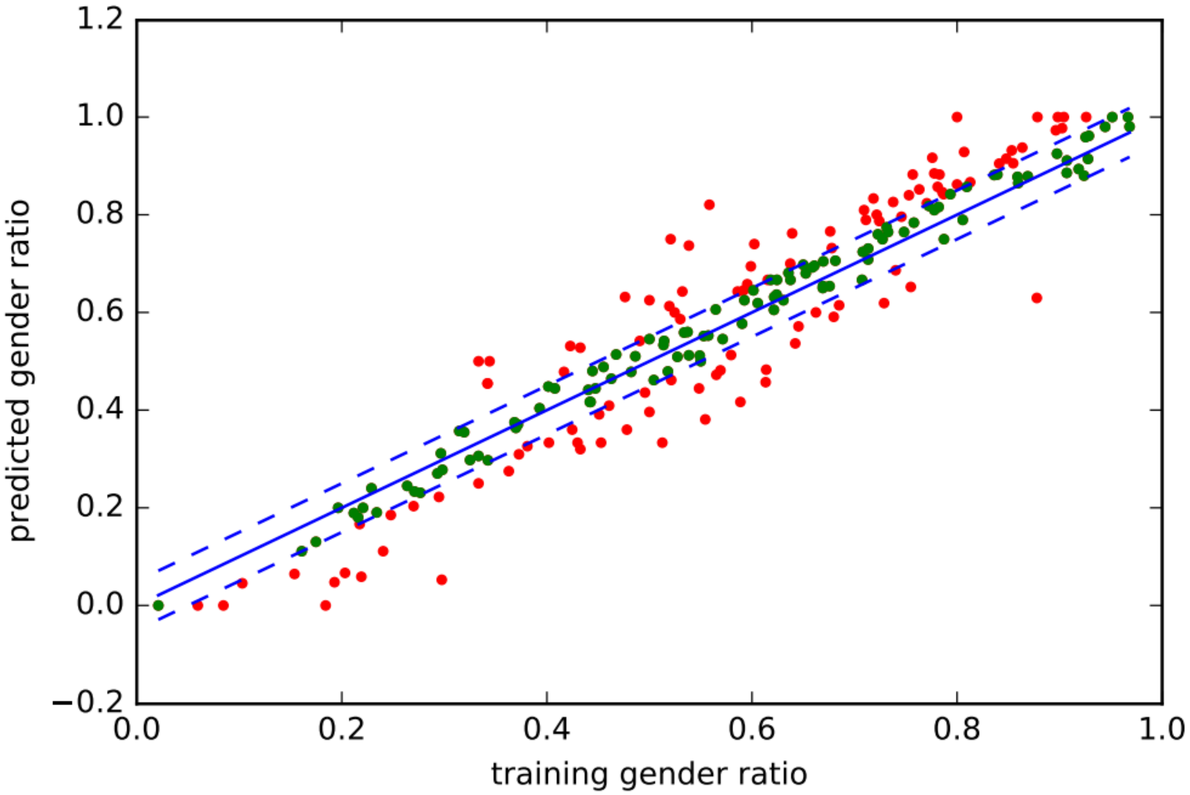
$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r\in M}}{\sum_i y^i_{v=v^*, r\in W} + \sum_i y^i_{v=v^*, r\in M}} \leq b^* + \gamma \quad (2)$$

Zhao et al. (2017)

Bias Amplification



(a) Bias analysis on imSitu vSRL without RBA



(c) Bias analysis on imSitu vSRL with RBA

- RBA: Reduce Bias Amplification

green: points meeting the margin

red: points violating the margin

Zhao et al. (2017)

Broad Areas

- Bias amplification: systems exacerbate real-world bias rather than correct for it
- Exclusion: underprivileged users are left behind by systems
- Dangers of automatic systems: automating things in ways we don't understand is dangerous
- Unethical use: powerful systems can be used for bad ends

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:

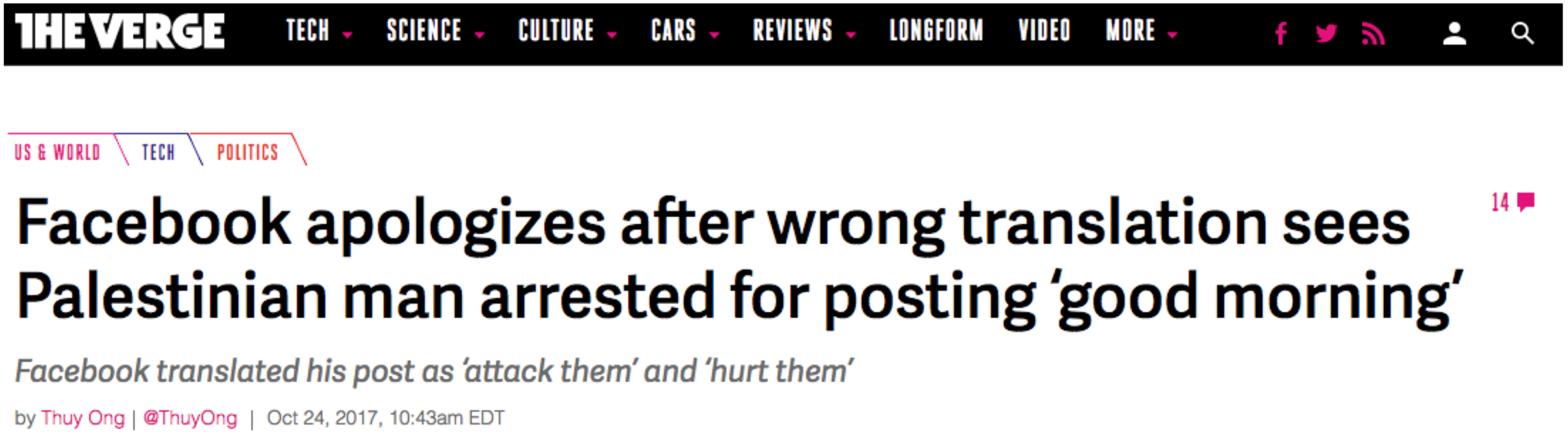
  Other dialects of English?

  Other languages? (Especially non-European/CJK)

  Codeswitching? (alternate between two or more languages)

- If important technological tools don't work for some users, where does that leave those users?

Dangers of Automatic Systems



Slide credit: The Verge

Dangers of Automatic Systems

- “Amazon scraps secret AI recruiting tool that showed bias against women”
  - “Women’s X” organization was a negative-weight feature in resumes
  - Women’s colleges too

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too
- Was this a bad model? May have actually modeled downstream outcomes correctly…but this can mean learning humans' biases

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

Charge-Based Prison Term Prediction with Deep Gating Network

Huajie Chen[1]* **Deng Cai**[2]* **Wei Dai**[1] **Zehui Dai**[1] **Yadong Ding**[1]
[1]NLP Group, Gridsum, Beijing, China
{chenhuajie,daiwei,daizehui,dingyadong}@gridsum.com
[2]The Chinese University of Hong Kong
thisisjcykcd@gmail.com

- Task: given case descriptions and charge set, predict the prison term

Case description: On July 7, 2017, when the defendant Cui XX was drinking in a bar, he came into conflict with Zhang XX…… After arriving at the police station, he refused to cooperate with the policeman and bited on the arm of the policeman……

Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*……

- Charge#1 creating disturbances term 12 months
- Charge#2 obstructing public affairs term 12 months

Chen et al. (EMNLP 2019)

Dangers of Automatic Systems

- Results: 60% of the time, the system is off by more than 20% (so 5 years => 4 or 6 years)

Model	S	EM	Acc@0.1	Acc@0.2
ATE-LSTM	66.49	7.72	16.12	33.89
MemNet	70.23	7.52	18.54	36.75
RAM	70.32	7.97	18.87	37.38
TNet	73.94	8.06	19.55	39.89
DGN	**76.48**	**8.92**	**20.66**	**42.61**

The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we recommend deploying our system in the "Review Phase", where other judges check the judgment result by a presiding judge. Our system can serve as one anonymous checker.

Dangers of Automatic Systems

- Results: 60% of the time, the system is off by more than 20% (so 5 years => 4 or 6 years)

Model	S	EM	Acc@0.1	Acc@0.2
ATE-LSTM	66.49	7.72	16.12	33.89
MemNet	70.23	7.52	18.54	36.75
RAM	70.32	7.97	18.87	37.38
TNet	73.94	8.06	19.55	39.89
DGN	**76.48**	**8.92**	**20.66**	**42.61**

- Is this the right way to apply this?
- Are there good applications this can have?
- Is this technology likely to be misused?

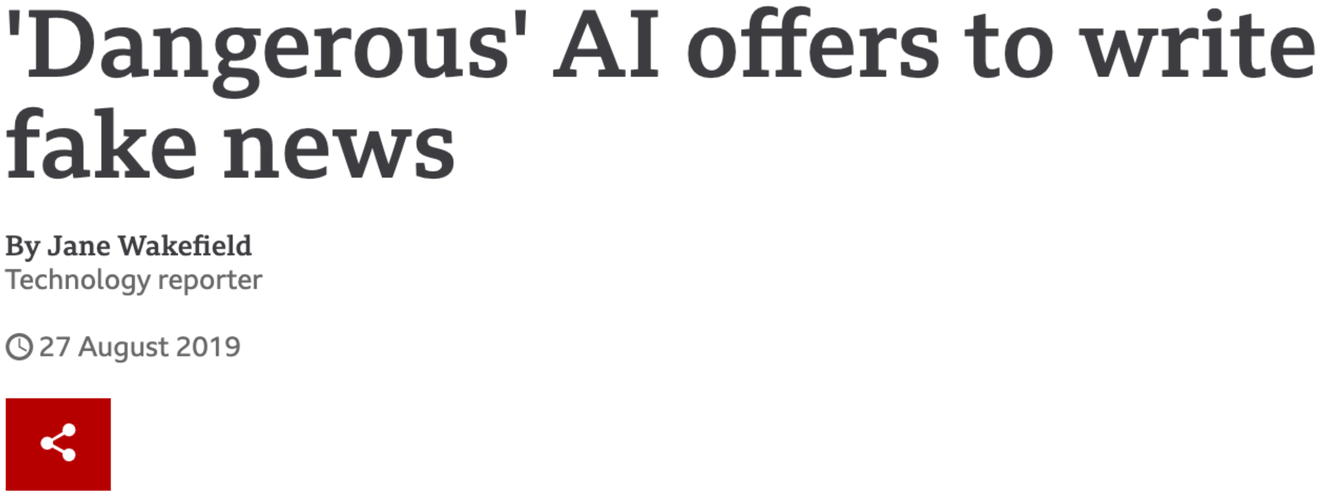
> The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we recommend deploying our system in the "Review Phase", where other judges check the judgment result by a presiding judge. Our system can serve as one anonymous checker.

Broad Areas

- Bias amplification: systems exacerbate real-world bias rather than correct for it
- Exclusion: underprivileged users are left behind by systems
- Dangers of automatic systems: automating things in ways we don't understand is dangerous
- Unethical use: powerful systems can be used for bad ends

Unethical Use

- Generating convincing fake news / fake comments?
  - What if these were undetectable?



'Dangerous' AI offers to write fake news

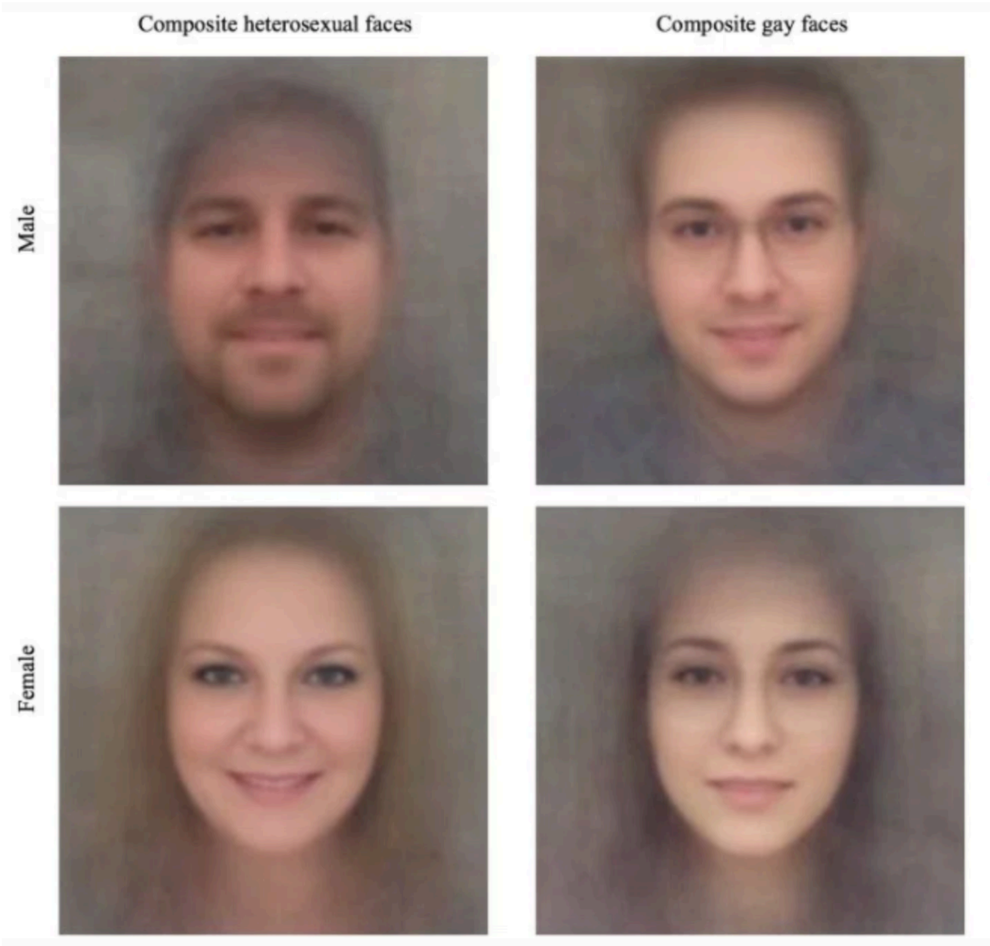
By Jane Wakefield
Technology reporter

27 August 2019

https://www.bbc.com/news/technology-49446729

Unethical Use

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: “this is useful because it supports a hypothesis” (physiognomy)
- Blog post by Agüera y Arcas, Todorov, Mitchell: mostly social phenomena (glasses, makeup, angle of camera, facial hair)
- If it’s not scientifically useful, the only ends might be bad ones



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

How to move forward

- Hal Daume III: Proposed code of ethics
  https://nlpers.blogspot.com/2016/12/should-nlp-and-ml-communities-have-code.html
  - Many other points, but these are relevant:
    - Contribute to society and human well-being, and minimize negative consequences of computing systems
    - Make reasonable effort to prevent misinterpretation of results
    - Make decisions consistent with safety, health, and welfare of public
    - Improve understanding of technology, its applications, and its potential consequences (pos and neg)
- Value-sensitive design: vsdesign.org
  - Account for human values in the design process: understand *whose* values matter here, analyze how technology impacts those values

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.
- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (though it’s not always easy to tell)
- As AI becomes more powerful, think about what we *should* be doing with it to improve society, not just what we *can* do with it

Administrivia

Reminder: SEI surveys

- HW3 graded (grades for grads will be uploaded later)
- Final project presentations on Dec 2 and 4.
  - See Carmen announcement to sign up
  - 10-minute presentation (including QA)
  - Can be “work in progress”, but should at least have preliminary results
  - Final reports due on December 6; no slip days
    - The format of your final report, e.g., https://arxiv.org/pdf/2010.12800.pdf

Have a relaxing Thanksgiving break!

Stay safe!!!