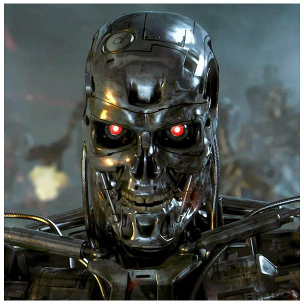
CS388: Natural Language Processing

Lecture 26: Wrapup and Ethics



Greg Durrett

TEXAS
The University of Texas at Austin

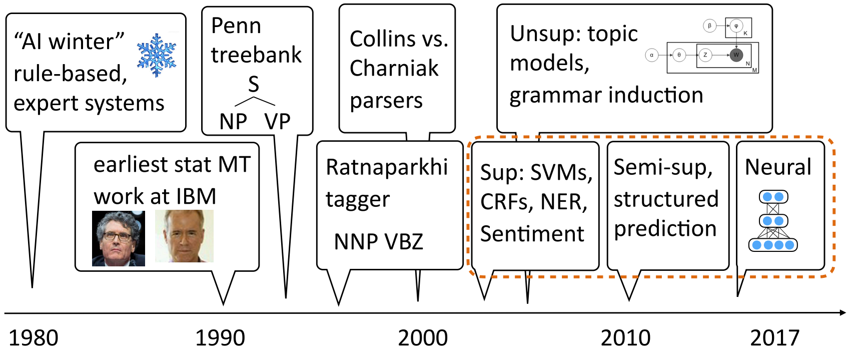
Administrivia

- My office hours are **today 2:30pm-3:30pm**. No OHs Weds/Thurs
- eCIS surveys
- Project 2 back
- Final project presentations next week
  - See Canvas announcement for who is presenting when
  - 3-minute presentations
  - Can be "work in progress", but should at least have preliminary results
  - Final reports due on December 13; no slip days

This Lecture

- Wrapup and current challenges
- Ethics in NLP/ML

A brief history of (modern) NLP



"AI winter" rule-based, expert systems

earliest stat MT work at IBM

Penn treebank

S

NP VP

Ratnaparkhi tagger

NNP VBZ

Collins vs. Charniak parsers

Unsup: topic models, grammar induction

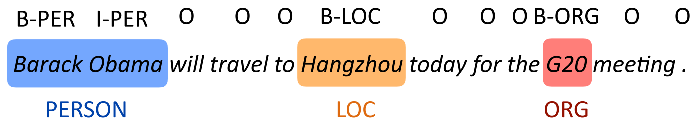
Sup: SVMs, CRFs, NER, Sentiment

Semi-sup, structured prediction

Neural

- What different model structures did we consider?

Sequential Structure: Analysis

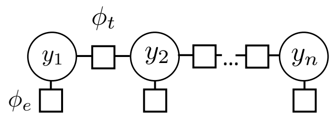


B-PER I-PER O O O B-LOC O O O B-ORG O O

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

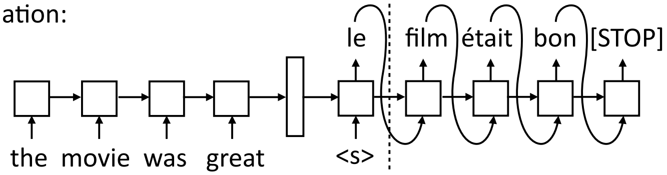
PERSON LOC ORG

- Can do language analysis with sequence models

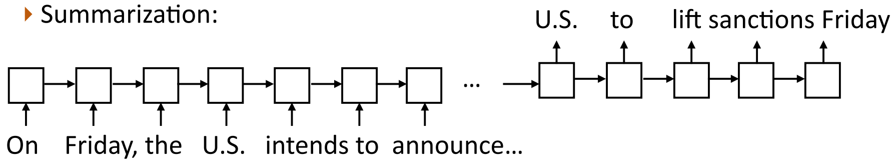


Sequential Structure: Generation

- Translation:

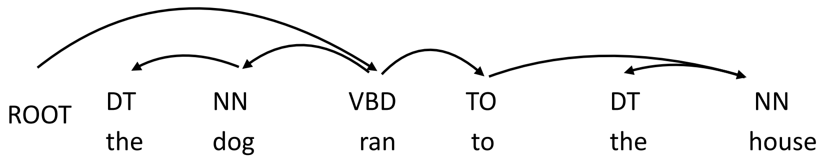


- Summarization:



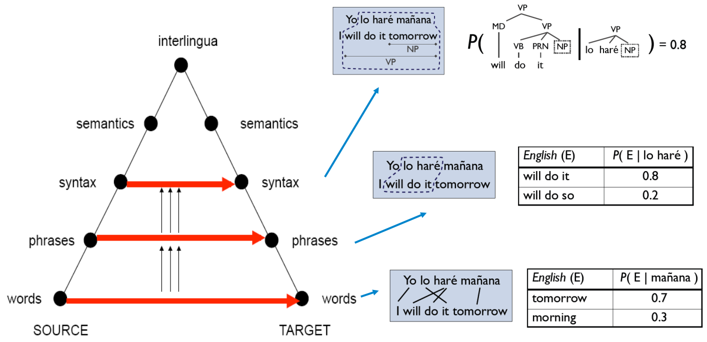
Tree Structure: Analysis

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



- Semantic roles: (ran, SUBJ=dog, IOBJ=house)
- AMRs that include coreference, etc.

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?

- 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?

- "Zero-shot" learning: test on a totally new domain

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

What can't we do?

- Deal with more "artistic" text. Dickens "A Tale of Two Cities":

It was the Dover road that lay, on a Friday night late in November, before the first of the persons with whom this history has business. The Dover road lay, as to him, beyond the Dover mail, as it lumbered up Shooter's Hill. He walked up hill in the mire by the side of the mail, as the rest of the passengers did; not because they had the least relish for walking exercise, under the circumstances, …

…

"Is that the Dover mail?"

"Why do you want to know?"

"I want a passenger, if it is."

"What passenger?"

"**Mr. Jarvis Lorry.**"

- Coreference? QA? Summarization? All extremely difficult

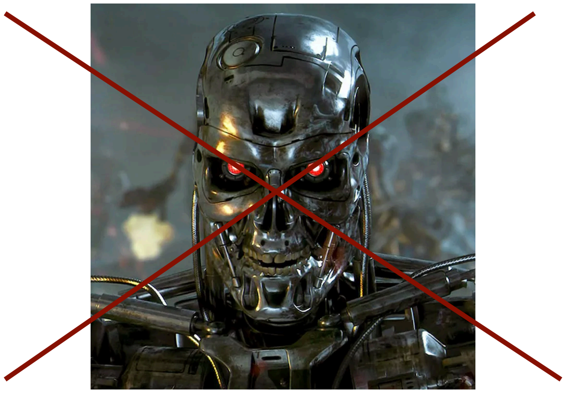
What can't we do?

- Commonsense

It was the Dover road that lay, on a Friday night late in November, before the first of the persons with whom this history has business. The Dover road lay, as to him, beyond the Dover mail, as it lumbered up Shooter's Hill. He walked up hill in the mire by the side of the mail, as the rest of the passengers did; not because they had the least relish for walking exercise, under the circumstances, …

  - Why are they walking by the mail?
  - Why wouldn't they relish the walking exercise?
- Reasoning about these sorts of scenarios is complicated, requires grounding and prior knowledge that current systems don't have

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
- 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 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?

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Zhao et al. (2017)

Bias Amplification

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

Maximize score of predictions…

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

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

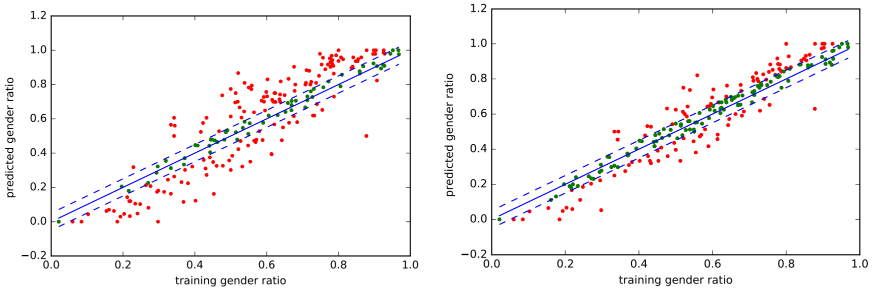
…subject to bias constraint

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

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

Zhao et al. (2017)

Bias Amplification

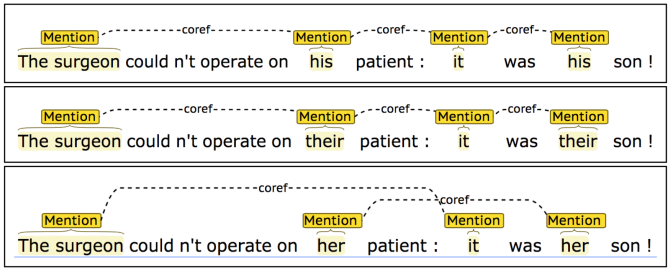


(a) Bias analysis on imSitu vSRL without RBA

(c) Bias analysis on imSitu vSRL with RBA

Zhao et al. (2017)

Bias Amplification



The surgeon could n't operate on his patient : it was his son !

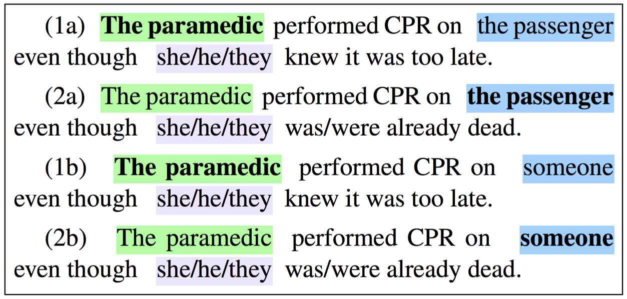
The surgeon could n't operate on their patient : it was their son !

The surgeon could n't operate on her patient : it was her son !

- Coreference: models make assumptions about genders and make mistakes as a result

Rudinger et al. (2018), Zhao et al. (2018)

Bias Amplification



(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

- Can form a targeted test set to investigate

Rudinger et al. (2018), Zhao et al. (2018)

Exclusion

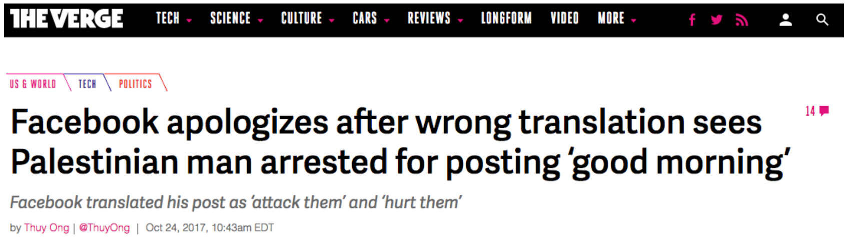
- Most of our annotated data is English data, especially newswire
- What about:

  Other dialects of English?

  Other languages? (Especially non-European/CJK)

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

Dangers of Automatic Systems



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

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

- "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

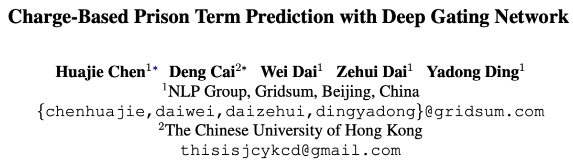
Dangers of Automatic Systems

"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

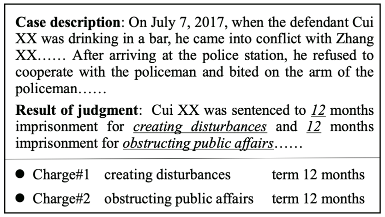
Dangers of Automatic Systems



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……

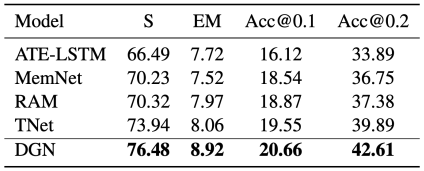
Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*……

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

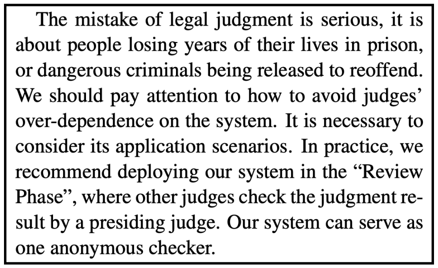
Chen et al. (EMNLP 2019)

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**

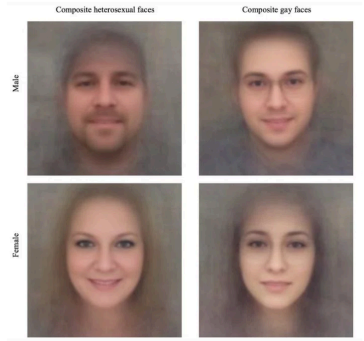


The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we 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.

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

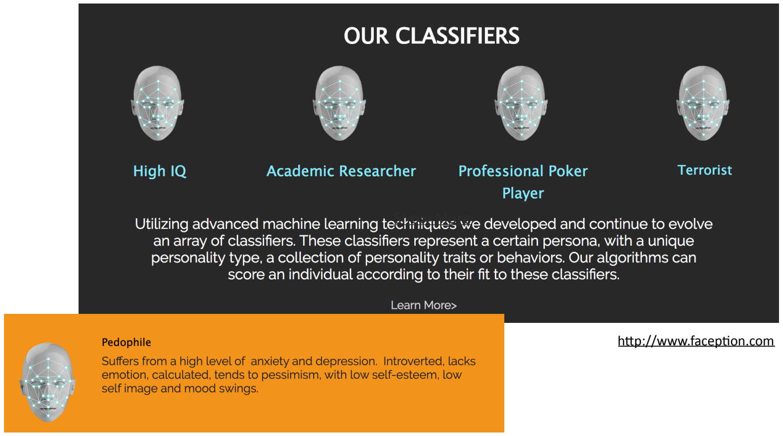
Unethical Use

- 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

Unethical Use



OUR CLASSIFIERS

Utilizing advanced machine learning techniques we developed and continue to evolve an array of classifiers. These classifiers represent a certain persona, with a unique personality type, a collection of personality traits or behaviors. Our algorithms can score an individual according to their fit to these classifiers.

Pedophile

Suffers from a high level of anxiety and depression. Introverted, lacks emotion, calculated, tends to pessimism, with low self-esteem, low self image and mood swings.

http://www.faception.com

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