A brief introduction to.
Natural Language Processing

Claire Cardie

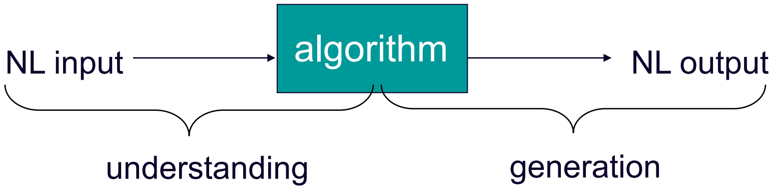
Professor

Computer Science and Information Science

Plan for the lecture

- What is NLP?
  - Levels of linguistic analysis
  - Issues that arise
- NLP applications
  - Focus: information extraction
  - Current trend

Natural Language Processing (NLP)


NL input

algorithm

NL output

understanding

generation

Linguistic analysis

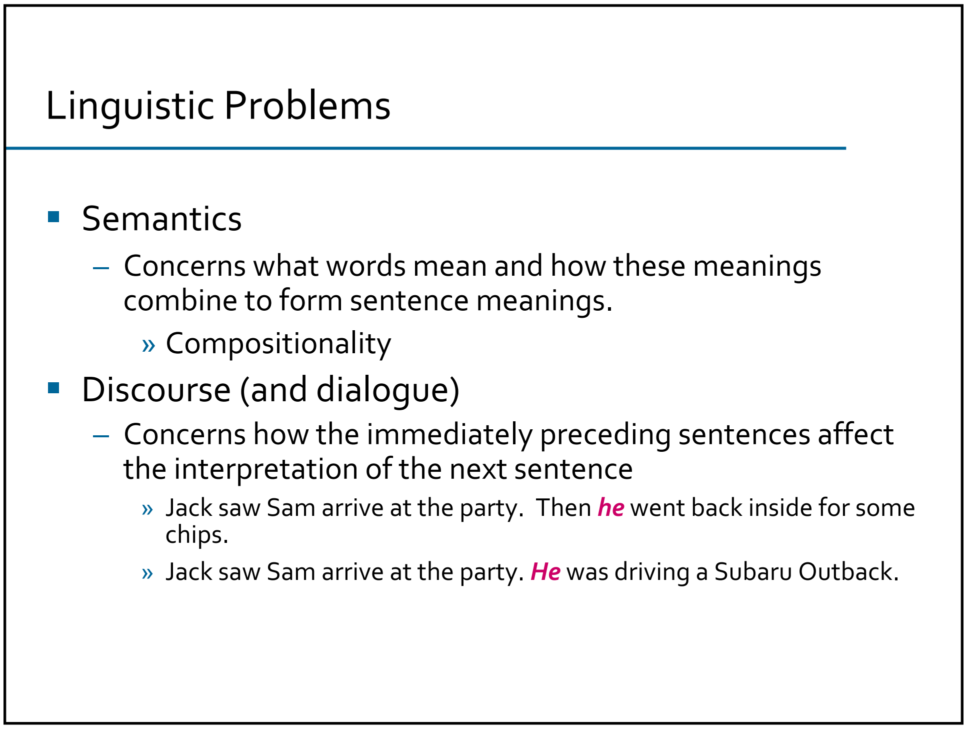
- Phonetics and phonology
  - Consider the phone [ni]
  - Corresponds to which word(s) in a corpus of phone conversations (AT&T)?
    - » knee
      - ◆ But this was the LOWEST probability option!!!!
    - » need
    - » neat
    - » new
      - ◆ Yeah, I live in New York.

Linguistic analysis

- Syntax
  - I ate spaghetti with meatballs.
  - I ate spaghetti with friends.
  - I ate spaghetti with a fork.
  - I ate spaghetti with glee.

Linguistic analysis

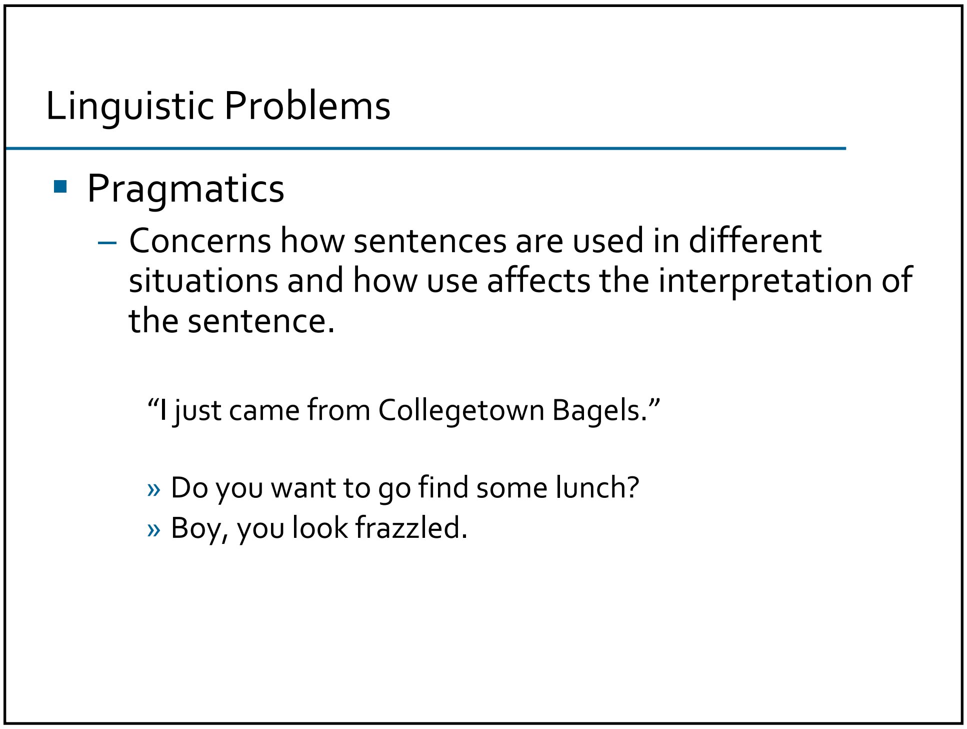
- Semantics
  - Lexical (word-level) semantics
    - Word-sense disambiguation
      - with → using
      - with → co-participant
      - …
      - Mirabelle drank from the **banks** of lake Cayuga.
        - Ground alongside a body of water?
        - Financial institution?



Linguistic Problems

- Semantics
  - Concerns what words mean and how these meanings combine to form sentence meanings.
    - Compositionality
- Discourse (and dialogue)
  - Concerns how the immediately preceding sentences affect the interpretation of the next sentence
    - Jack saw Sam arrive at the party. Then ***he*** went back inside for some chips.
    - Jack saw Sam arrive at the party. ***He*** was driving a Subaru Outback.

7



Linguistic Problems

- Pragmatics
  - Concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

    "I just came from Collegetown Bagels."

    - Do you want to go find some lunch?
    - Boy, you look frazzled.

8

Key issue

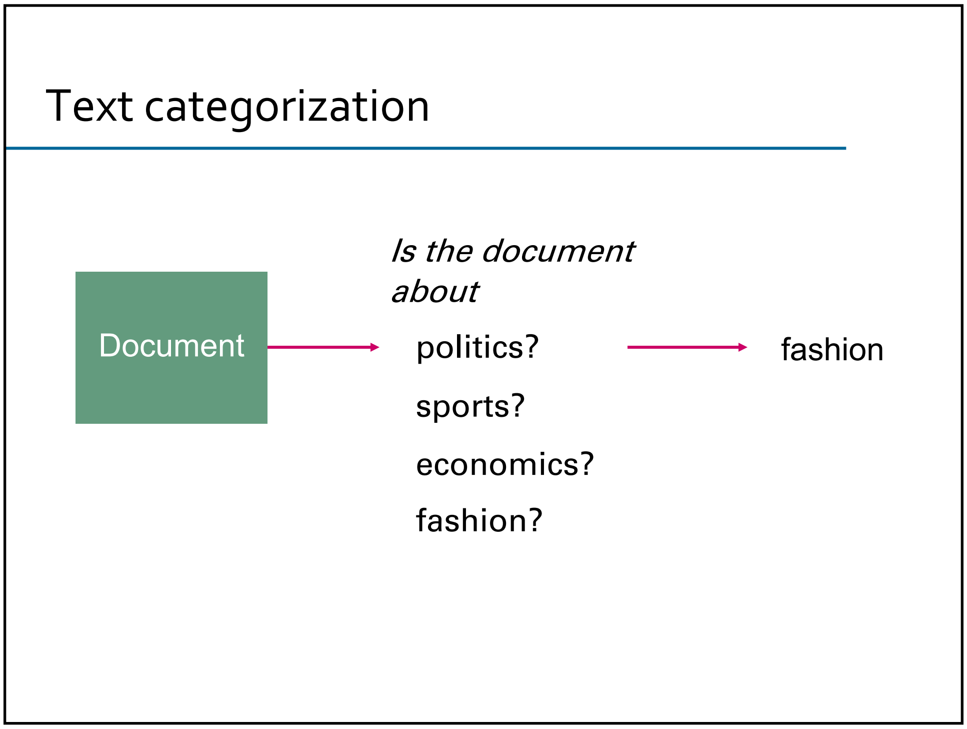
- Ambiguity!!!!
  - At all levels of linguistic analysis
- Difficult (impossible) to design algorithms based on our intuitions
- Solution:
  - Rely on machine learning methods

9

Plan for the lecture

- What is NLP?
  - Levels of linguistic analysis
  - Issues that arise
- NLP applications
  - Focus: information extraction
  - Current trend

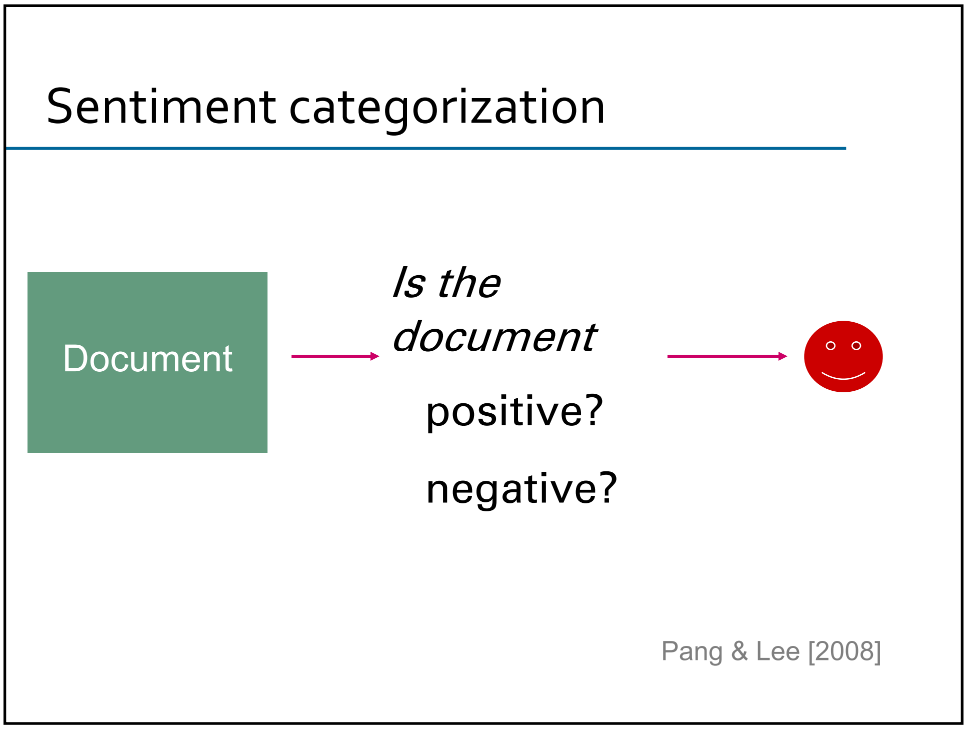
10



Text categorization

Document → *Is the document about* politics? sports? economics? fashion? → fashion

11



Sentiment categorization

Document → *Is the document* positive? negative? → ☺

Pang & Lee [2008]

12

Summarization

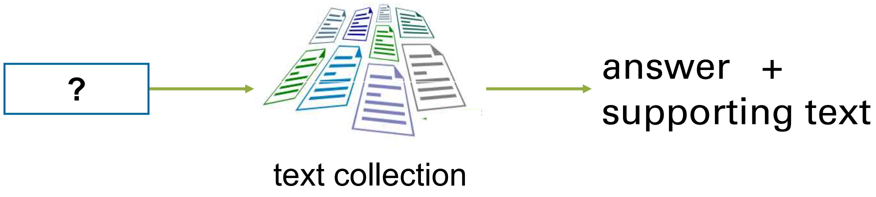
- headlines (from around the world)
- outlines (notes for students)
- minutes (of a meeting)
- previews (of movies)
- synopses (soap opera listings)
- reviews (of a book, CD, movie, etc.)
- digests (TV guide)
- biography (resumes, obituaries)
- abridgments (Shakespeare for children)
- bulletins (weather forecasts/stock market reports)
- sound bites (politicians on a current issue)
- histories (chronologies of salient events)

involves
natual language generation!!!

13

Question answering

- Task
  - » How many calories are there in a Big Mac?
  - » Who is the voice of Miss Piggy?
  - » Who was the first American in space?
  - – Retrieve not just relevant documents, but return the answer



15

Machine translation

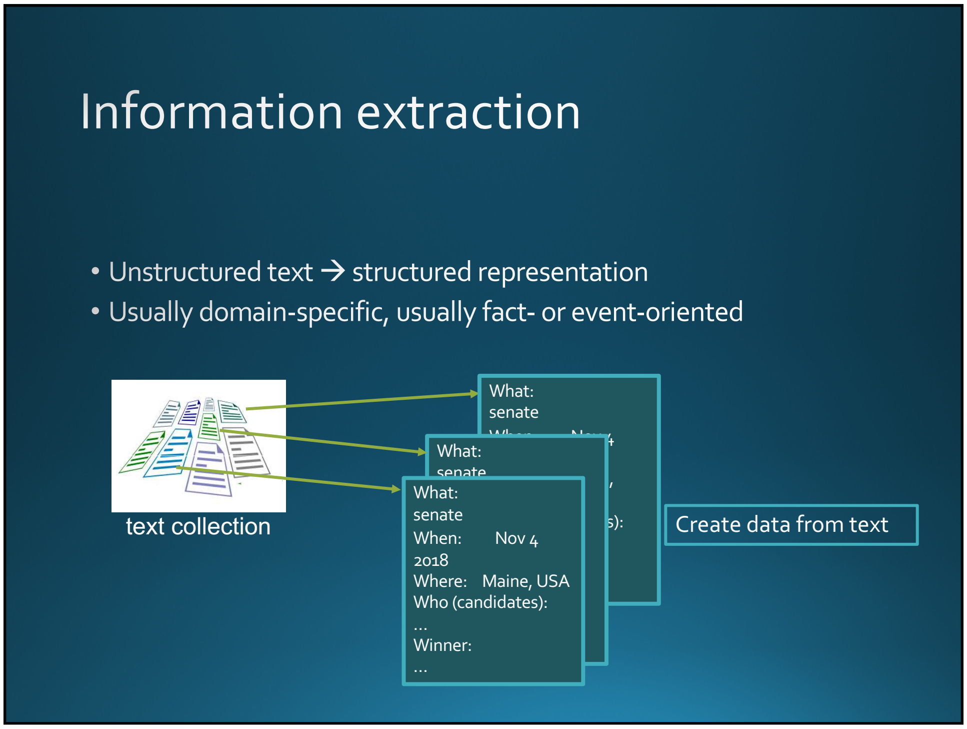
- one of the first applications envisioned for NLP techniques
  - *vodka good meat rotten*
    The ~~spirit~~ is ~~willing~~, but the ~~flesh~~ is ~~weak~~.
  - "open"

17

Focus on one application

- Information extraction

19



Information extraction

- Unstructured text → structured representation
- Usually domain-specific, usually fact- or event-oriented

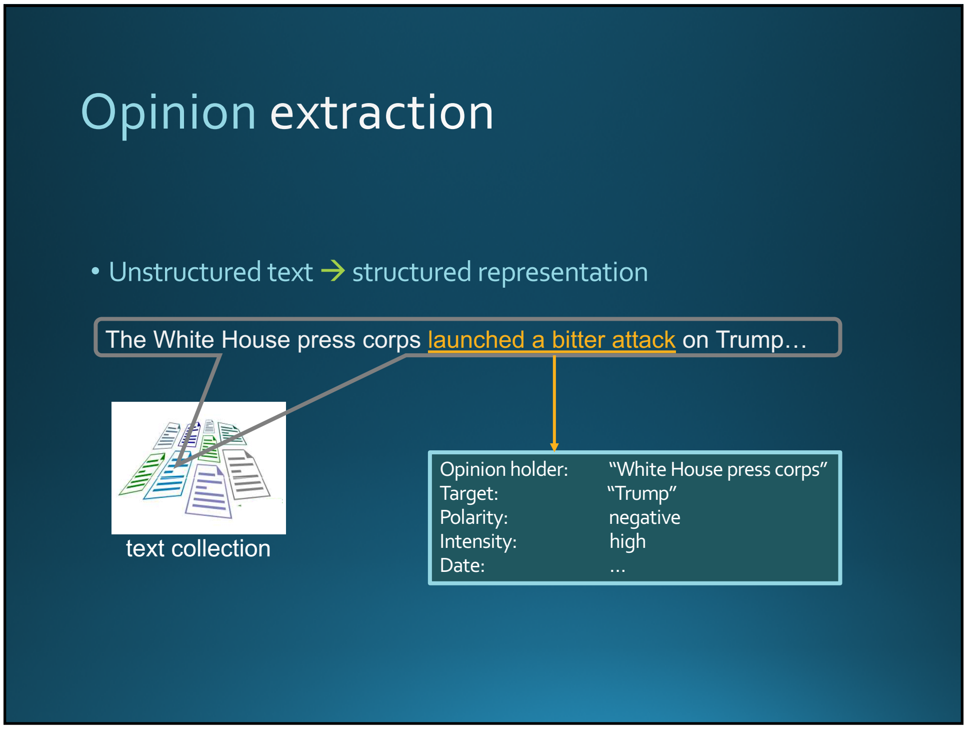
What:
senate

What:
senate

What:
senate
When: Nov 4 2018
Where: Maine, USA
Who (candidates):
…
Winner:
…

Create data from text

20



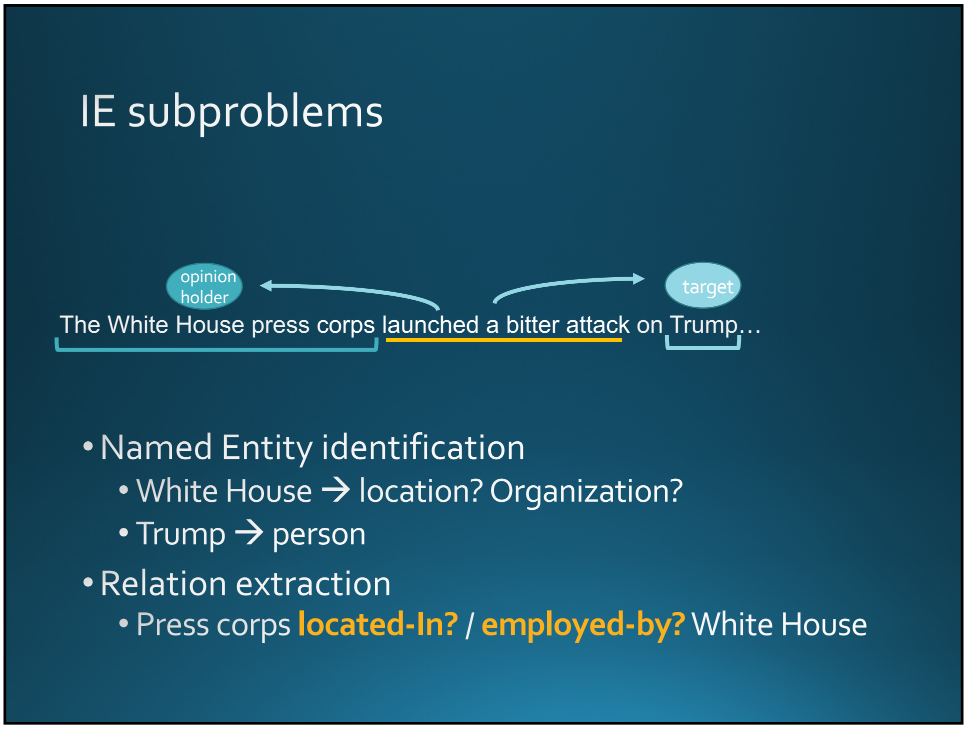
Opinion extraction

- Unstructured text → structured representation

The White House press corps launched a bitter attack on Trump…

Opinion holder:	"White House press corps"
Target:	"Trump"
Polarity:	negative
Intensity:	high
Date:	…

21



IE subproblems

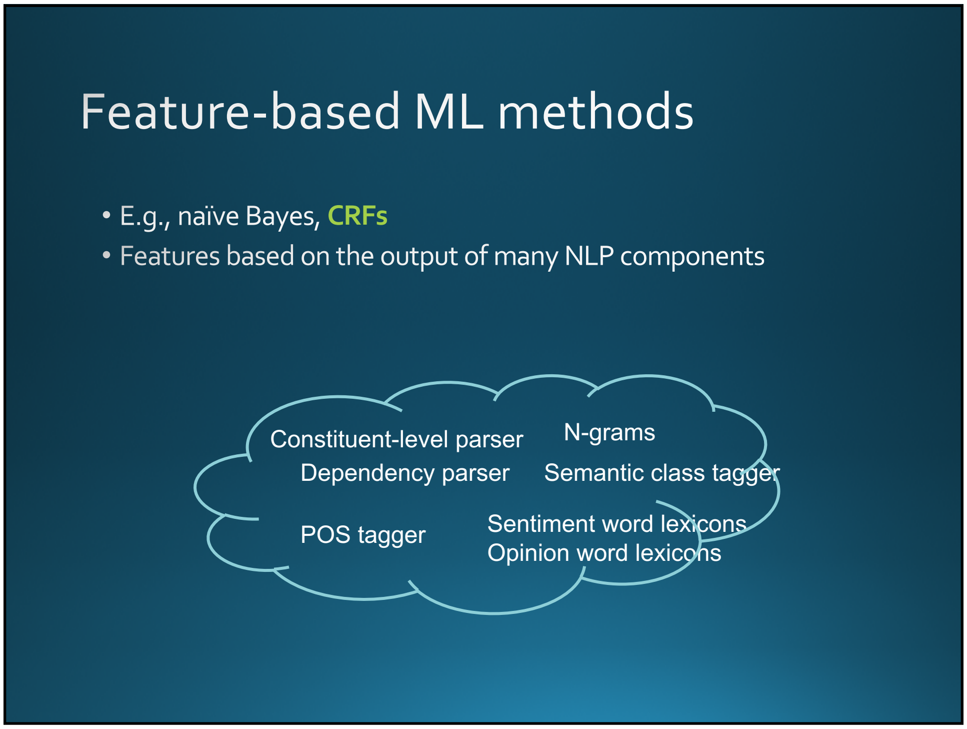
opinion holder

target

The White House press corps launched a bitter attack on Trump…

- Named Entity identification
  - White House → location? Organization?
  - Trump → person
- Relation extraction
  - Press corps **located-In?** / **employed-by?** White House

22



Feature-based ML methods

- E.g., naïve Bayes, **CRFs**
- Features based on the output of many NLP components

Constituent-level parser

N-grams

Dependency parser

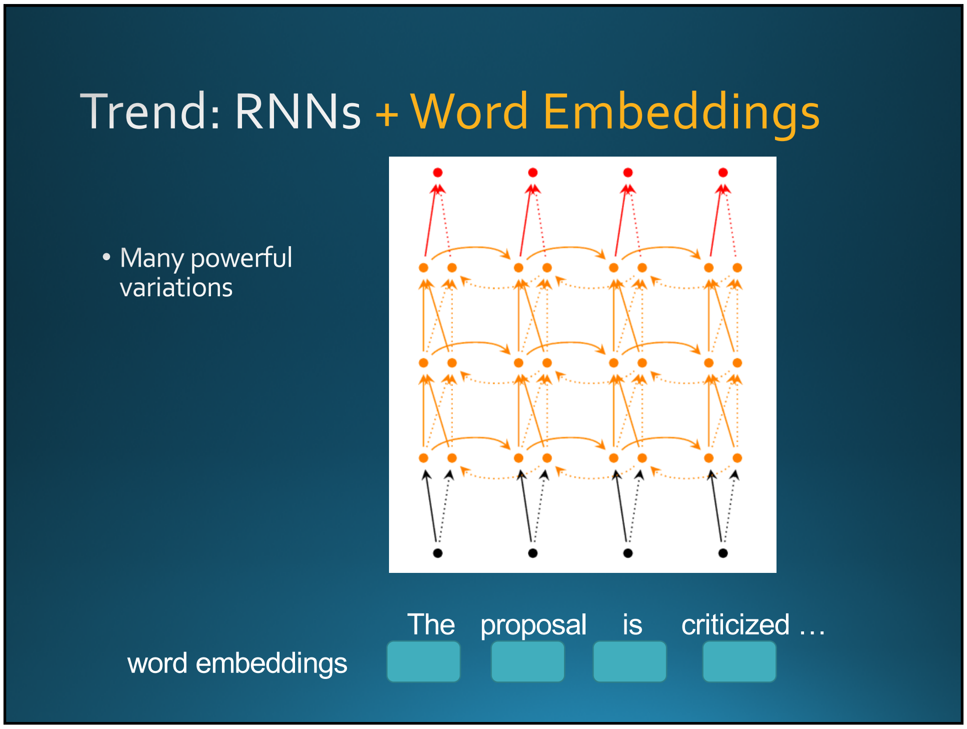
Semantic class tagger

POS tagger

Sentiment word lexicons

Opinion word lexicons

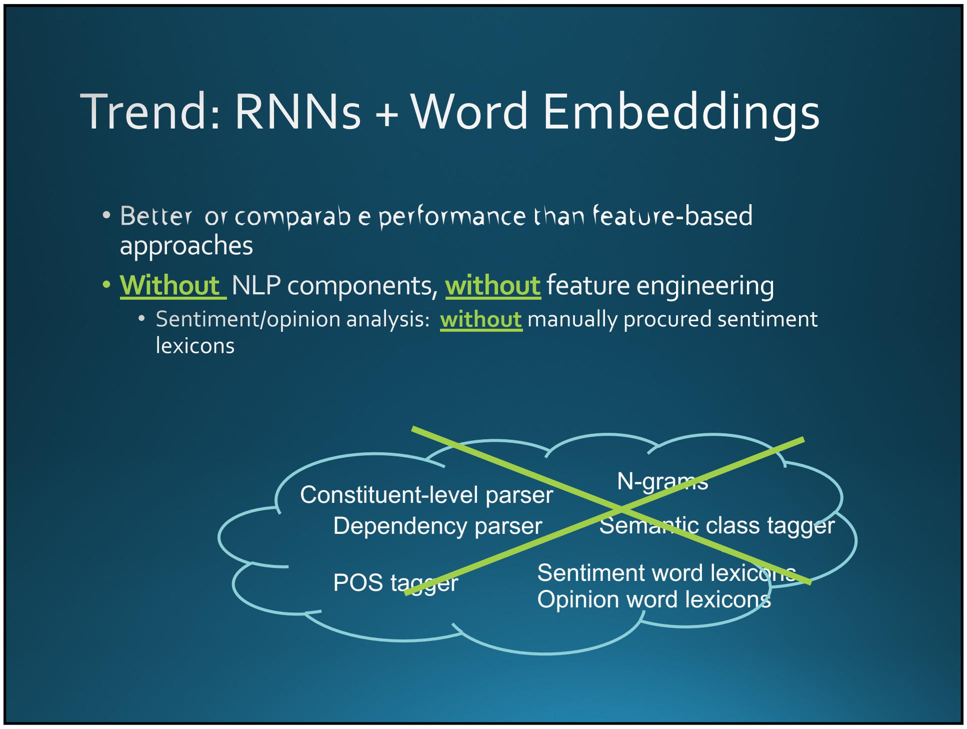
23



Trend: RNNs + Word Embeddings

- Many powerful variations

24



Trend: RNNs + Word Embeddings

- Better or comparable performance than feature-based approaches
- **Without** NLP components, **without** feature engineering
  - Sentiment/opinion analysis: **without** manually procured sentiment lexicons

~~Constituent-level parser~~
~~Dependency parser~~
~~POS tagger~~
~~N-grams~~
~~Semantic class tagger~~
~~Sentiment word lexicons~~
~~Opinion word lexicons~~

25

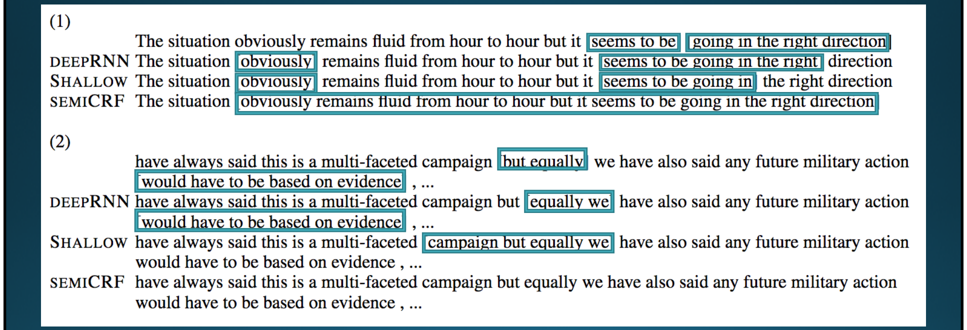
Opinion analysis

opinion expressions	proportional overlap	exact match
CRF	64.4 F	57.7 F
CRF + word embeddings	66.4 F	59.6 F
Deep bidirectional RNNs	71.7 F	66.0 F

[Irsoy & Cardie, EMNLP2014]

26

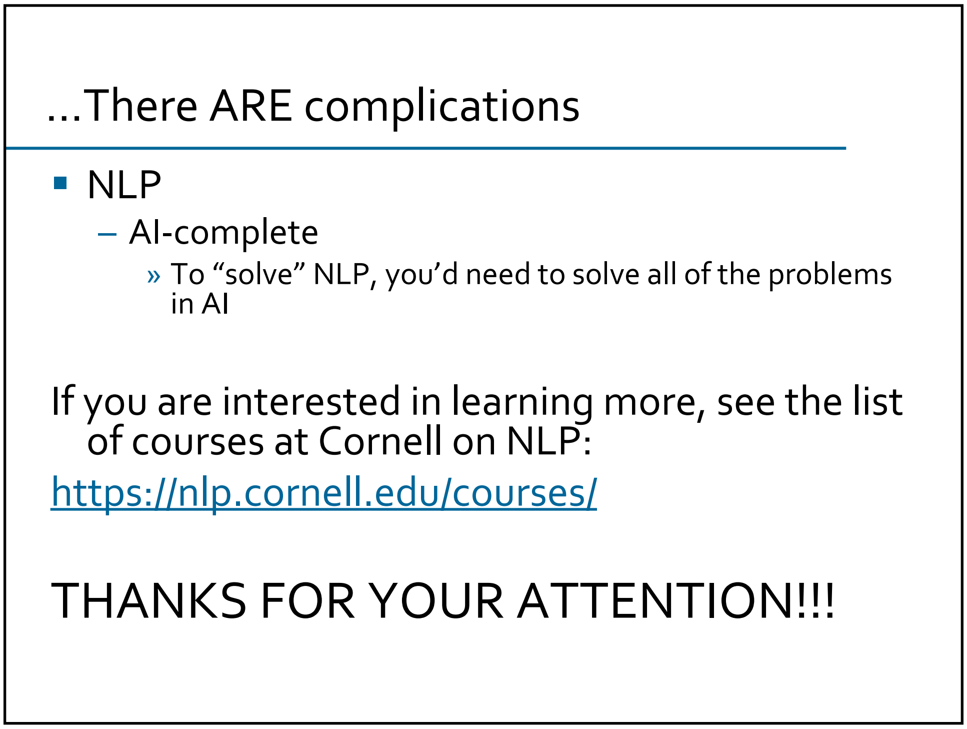
Opinion expressions



(1)
The situation obviously remains fluid from hour to hour but it seems to be going in the right direction
DEEPRNN The situation obviously remains fluid from hour to hour but it seems to be going in the right direction
SHALLOW The situation obviously remains fluid from hour to hour but it seems to be going in the right direction
SEMICRF The situation obviously remains fluid from hour to hour but it seems to be going in the right direction

(2)
have always said this is a multi-faceted campaign but equally we have also said any future military action would have to be based on evidence , ...
DEEPRNN have always said this is a multi-faceted campaign but equally we have also said any future military action would have to be based on evidence , ...
SHALLOW have always said this is a multi-faceted campaign but equally we have also said any future military action would have to be based on evidence , ...
SEMICRF have always said this is a multi-faceted campaign but equally we have also said any future military action would have to be based on evidence , ...

27



…There ARE complications

- NLP
  - AI-complete
    - To “solve” NLP, you’d need to solve all of the problems in AI

If you are interested in learning more, see the list of courses at Cornell on NLP:

https://nlp.cornell.edu/courses/

THANKS FOR YOUR ATTENTION!!!

29