

CS-AD 220 – Spring 2016

Natural Language Processing

Session 2: 2-Feb-16

Prof. Nizar Habash

Some slides are adapted from Jurafsky and Martin's course
slides on Speech and Language Processing

Roadmap

- Natural Language Processing
 - Knowledge
 - Ambiguity
 - Models and Algorithms
 - Challenges
 - State of the art

Linguistic Knowledge for NLP

- Consider the interaction with HAL the computer from 2001: A Space Odyssey
 - Dave: *Open the pod bay doors, Hal.*
 - HAL: *I 'm sorry Dave, I 'm afraid I can 't do that.*
- What applications are needed?
- What kinds of knowledge is needed?

What's needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
 - How groups of words clump
 - What the words and the clumps mean
- Dialog Systems
 - It is polite to respond, even if you're planning to kill someone.
 - It is polite to pretend to want to be cooperative (I'm afraid, I can't...)

Caveat

- NLP has an AI aspect to it.
 - We're often dealing with ill-defined problems
 - We don't often come up with exact solutions/algorithms
 - We can't let either of those facts get in the way of making progress
- Turing Test

Turing Test

- Alan Turing was British pioneering computer scientist, mathematician, logician, and cryptanalyst. He is widely considered the Father of Computer Science.
- The movie *Imitation Game* is about him.
- The Turing test is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human. Turing proposed that a human evaluator would judge natural language conversations between a human and a machine that is designed to generate human-like responses.



Categories of Linguistic Knowledge

- Phonology
 - Morphology
 - Syntax
 - Semantics
 - Pragmatics
 - Discourse
- Each kind of knowledge has associated with it an encapsulated set of processes that make use of it.
 - Interfaces are defined that allow the various levels to communicate.
 - This usually leads to a pipeline architecture.

Consider...

“I’m afraid I can’t do that.”

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

Ambiguity

- Find at least 5 meanings of this sentence:
 - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence:
 - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
 - **Lexical category**: “duck” can be a Noun or a Verb
- I cooked waterfowl belonging to her.
 - **Lexical category**: “her” can be a possessive (“of her”) or dative (“for her”) pronoun
- I made the (plaster) duck statue she owns
 - **Lexical Semantics**: “make” can mean “create” or “cook”

Ambiguity is Pervasive

- **Grammar:** *make* can be...
 - **Transitive: (verb has a noun direct object)**
 - I cooked [waterfowl belonging to her]
 - **Ditransitive: (verb has 2 noun objects)**
 - I made [her] (into) [undifferentiated waterfowl]
 - **Action-transitive (verb has a direct object and another verb)**
 - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
 - I mate or duck
 - I'm eight or duck
 - Eye maid; her duck
 - Aye mate, her duck
 - I maid her duck
 - I'm aid her duck
 - I mate her duck
 - I'm ate her duck
 - I'm ate or duck
 - I mate or duck

Ambiguity is pervasive

I saw the man with a telescope

I saw the man with a telescope

Ambiguity is pervasive

New York Times headline (17 May 2000)

Fed raises interest rates

Fed raises interest rates

Fed raises interest rates 0.5%

Analysis vs. Disambiguation

بين



PV+PVSUFF_SUBJ:3MS	bay~an+a	He demonstrated
PV+PVSUFF_SUBJ:3FP	bay~an+~a	They demonstrated (f.p)
NOUN_PROP	biyn	Ben
ADJ	bay~in	Clear
PREP	bayn	Between, among

Morphological Analysis
Morphological Disambiguation

is out-of-context
is in-context

Analysis vs. Disambiguation

Will Ben Affleck be a good Batman?

هل سينجح بين أفليك في دور باتمان؟



PV+PVSUFF_SUBJ:3MS	bay~an+a	He demonstrated
PV+PVSUFF_SUBJ:3FP	bay~an+~a	They demonstrated (f.p)
* NOUN_PROP	biyn	Ben
ADJ	bay~in	Clear
PREP	bayn	Between, among

Morphological Analysis

is out-of-context

Morphological Disambiguation

is in-context

Dealing with Ambiguity

Four possible approaches

1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide at ambiguous levels.
2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.
3. Probabilistic approaches based on making the most likely choices.
4. Don't do anything, maybe it won't matter.
 1. *We'll leave when the duck is ready to eat.*
 2. *The duck is ready to eat now.*
 - Does the “duck” ambiguity matter with respect to whether we can leave?

Models and Algorithms

- By **models** we mean the formalisms that are used to capture the various kinds of **linguistic knowledge** we need.
- **Algorithms** are then used to manipulate the **knowledge representations** needed to tackle the task at hand.

Models

- State machines
- Rule-based approaches
- Logical formalisms
- Probabilistic models

Algorithms

- Many of the algorithms that we'll study will turn out to be **transducers**; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds

Paradigms

- In particular..
 - State-space search
 - To manage the problem of making choices during processing when we lack the information needed to make the right choice
 - Dynamic programming
 - To avoid having to redo work during the course of a state-space search
 - CKY, Earley, Minimum Edit Distance, Viterbi, etc.
 - Classifiers
 - Machine learning based classifiers that are trained to make decisions based on features extracted from the local context

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse

get cold feet

lose face

throw in the towel

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse

get cold feet

lose face

throw in the towel

neologisms

unfriend

Retweet

bromance

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing ...
Let It Be sold millions...
... a mutation on the *for* gene ...

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing ...
Let It Be sold millions...
... a mutation on the *for* gene ...

But that's what makes it fun!

Making progress on this problem...

- The task is difficult! What tools do we need?
 - Knowledge about language
 - Knowledge about the world
 - A way to combine knowledge sources
- How we generally do this:
 - Probabilistic models built from language data
 - $P(\text{"maison"} \rightarrow \text{"house"})$ **high**
 - $P(\text{"L'avocat général"} \rightarrow \text{"the general avocado"})$ **low**
 - Luckily, rough text features can often do half the job.

Language Technology

making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.


Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Sentiment analysis

Best roast chicken in San Francisco! 

The waiter ignored us for 20 minutes. 

Coreference resolution

Carter told Mubarak he shouldn't run again.



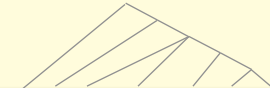
Word sense disambiguation (WSD)

I need new batteries for my *mouse*.



Parsing

I can see Alcatraz from the window!



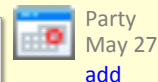
Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose



Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?



Adapted from Speech and Language Processing - Jurafsky and Martin

Real Success: IBM's Watson

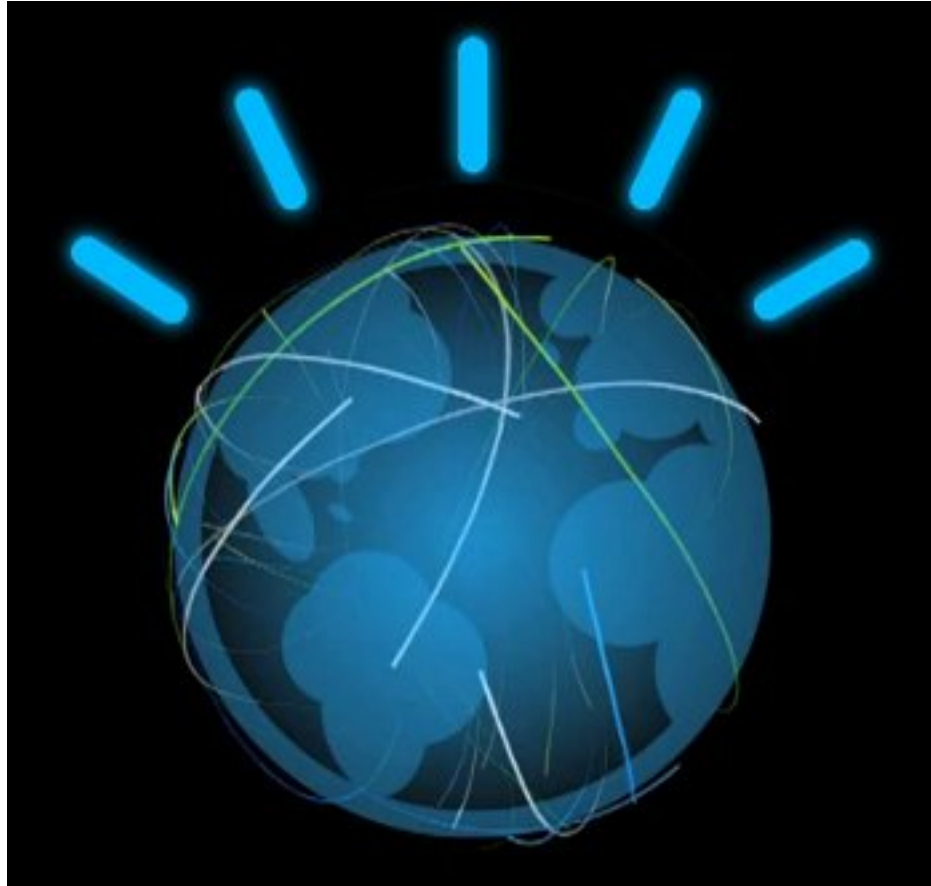
- Won Jeopardy on February 16, 2011!

**WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL**



Bram Stoker

Real: Watson on Jeopardy



- https://www.youtube.com/watch?v=WFR3lOm_xhE

Next Session

- Basic text processing
 - Unix tools
 - Regular expressions
- Bring your laptop to class!