Human Language Technologies: CS 1671, Fall 2018

Course Overview

&

Introduction (Chapter 1)

Professor Litman

- Bachelors in CS/Math from William and Mary
- PhD in CS from University of Rochester
  - Dissertation in human language technologies
- Many years at AT&T Bell Laboratories
  - Artificial Intelligence Department
  - Stint as Assistant Professor at Columbia University
- Joined Pitt in 2001
  - Professor, Computer Science
  - Co-Director, Intelligent Systems Program
  - Senior Scientist, Learning Research & Development Center
  - See homepage for research
- Fellow, Association for Computational Linguistics

2

Human Language Technologies (HLT)

- Also known as
  - Natural Language Processing (NLP)
  - Computational Linguistics (CL)
- Intersects with
  - Artificial Intelligence (AI)
  - Machine Learning (ML)
- Increasing interaction with
  - Computer Vision

3

Natural Language Processing

- We're going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
- We are less concerned with the insights that such computational work gives us into human processing of language (cognitive science, psycholinguistics)

Why Should You Care?

1. An enormous amount of knowledge is now available in machine readable form as natural language text, audio, and video
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Commercial World

- Lots of exciting stuff going on, e.g.

Google Translate

Google Translate

Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Question Answering to Jeopardy Winner

Web Analytics

- Text-mining of blogs, twitter, discussion forums, message boards, user groups, and other forms of social media
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what's hot, what topics are people talking about right now).

Personal Assistants

11

What is NLP?

- Natural Language (**NL**P)
  - Chinese, English, Spanish, Hindi, etc.
- Processing (NL**P**)
  - analysis (NL -> R)
  - generation (R -> NL)
  - acquisition of R from knowledge and data
- What is R?
  - A representation, typically depends on the application

12

Example NLP Desiderata

- Sensitivity to a wide range of the phenomena and constraints in human languages
- Generality across different languages, genres, styles, and modalities
- Computational efficiency at construction time and runtime
- Strong theoretical guarantees (e.g., convergence)
- High accuracy when judged against expert performance

13

Major Topics

1. Words
2. Syntax
3. Semantics (Meaning)
4. Discourse & Dialog(ue)

5. Applications exploiting each

Applications

- First, what makes an application a *language processing application* (as opposed to any other piece of software)?
  - An application that requires the use of knowledge about human languages
    - Example: Is Unix wc (word count) an example of a language processing application?

Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities

Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the following interaction with HAL the computer from 2001: A Space Odyssey

HAL from 2001

- Dave: *Open the pod bay doors, Hal.*
- HAL: *I'm sorry Dave, I'm afraid I can't do that.*

What's needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
- How groups of words clump
  - What the clumps mean

What's needed?

- Dialog
  - 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

Course Material

- We'll be intermingling discussions of:
  - Linguistic topics
    - E.g. Morphology, syntax, discourse structure
  - Formal systems
    - E.g. Regular languages, context-free grammars
  - Applications
    - E.g. Chatbots, information extraction

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical, sentence, document semantics
- Discourse processing
- Dialogue structure

Topics: Techniques

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Lexical
- (First order logic)

- Probability models
- Supervised machine learning methods

Topics: Applications

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Stand-alone
- Enabling applications
- Funding/Business plans

Categories of 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.

Morphology

- Friend
- Friends
- Unfriend
- Friendly

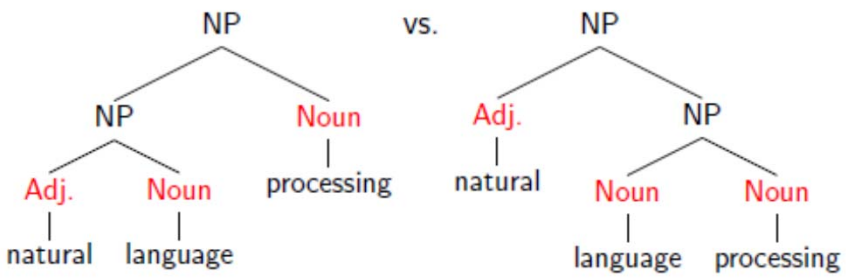
- Important for search

The Challenges of "Words"

- Segmenting text into words (e.g., Thai)
- Morphological variation (e.g. Turkish)
- Words with multiple meaning (e.g., bank, latex)
- Multiword expressions (e.g., Squirrel Hill)
- Non-standard forms (e.g. u -> you)

29

Syntax



30

- Is this a syntactically valid sentence of English?
  - Buffalo buffalo buffalo Buffalo buffalo.

31

Syntax + Semantics

- We saw the man with the telescope wrapped in paper.
  - Who has the telescope?
  - Who or what is wrapped in paper?
  - An event of perception or an assault?

32

Semantics

- Every fifteen minutes a woman in this country gives birth.
- Our job is to find this woman, and stop her!
  - Groucho Marx

33

Ambiguity

- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

Ambiguity in Headlines

- Government Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half

35

Subtler Ambiguity

Q: Why does my high school give me a suspension for skipping class?

A: Administrative error. They're supposed to give you a suspension for auto shop, and a jump rope for skipping class.

36

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 N or V
- 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

Dealing with Ambiguity

- Some possible approaches:
  1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide among choices at ambiguous levels.
  2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.

Dealing with Ambiguity

3. Probabilistic approaches based on making the most likely choices
4. Don't do anything, maybe it won't matter

More difficulties: Variation and noise in linguistic input

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

The Language of Genres: news, journal, social media, novel?

- Devasting, but yet amazing storm. IMBY just some branches, ton of leaves, ..
- Produced by a team of 26 scientists led by the University of New South Wales Climate Research Centre, the Diagnosis convincingly proves that ...
- Hurricane Sandy churned about 290 miles off the MidAtlanic coast Sunday night, with the National Hurricane Center reporting...
- What is the matter?' I cried. 'A wreck! Close by!'

45

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

- Regular Expressions
- Rule-based approaches
- Logical formalisms
- Vector/neural models
- Probabilistic models

Algorithms

- Many of the algorithms that we'll study 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

- For example..
  - 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 a state-space search
      - Minimum Edit Distance, Viterbi, CKY
  - Classifiers
    - Machine learning based classifiers that are trained to make decisions based on features extracted from the local context
  - Alignment

State Space Search

- States represent pairings of partially processed inputs with partially constructed representations.
- Goals are inputs paired with completed representations that satisfy some criteria.
- As with most interesting problems the spaces are normally too large to exhaustively explore.
  - We need heuristics to guide the search
  - Criteria to trim the space

Dynamic Programming

- Don't do the same work over and over.
- Avoid this by building and making use of solutions to sub-problems that must be invariant across all parts of the space.

NLP & Machine Learning

- ML needs bias/assumptions and features, which might be linguistic
- NLP provides inspiring applications for ML

NLP and Linguistics

- NLP must contend with data as found in the real world
- “Every time I fire a linguist, the performance of our speech recognition system goes up.”

53

Some Factors Changing NLP

- Computing power
- The web
- Social media
- Deep learning
- Language in social context
  - Hirschberg and Manning, 2015

54

Next Class


WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL


YOU KNOW HOW SOMETIMES PEOPLE PUT A SPACE IN THEIR EMAIL ADDRESS TO MAKE IT HARDER TO HARVEST?
YEAH?
THEY HAVE A TOOL THAT CAN DELETE THE SPACE!
OH MY GOD.

5