COSC 6336
Natural Language Processing

Instructor: Thamar Solorio

Some content in these slides was adapted from Prof. Rada Mihalcea

Natural?

Natural Language?

language spoken by people, e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc.

Natural Language Processing

Applications that deal with natural language in a way or another

Computational Linguistics

Doing linguistics on computers, more on the linguistic side than NLP, but closely related

Why Natural Language Processing?

- ★ " … language is what made us human" (Guy Deutscher)
- ★ Through language humans:
  - Pass on knowledge
  - Create new thoughts and ideas
  - Express deep (and not so deep) reflections
- ★ Practical value:
  - Companies want to know what consumers are saying
  - Intelligence communities want to know what persons of interest are planning
  - New products that use language as the interface with humans
- ★ Scientific value:
  - Gain a deeper understanding of how the human brain is able to process language

Why Natural Language Processing?

- We are constantly generating data
- Computer programs that can process this data require NLP expertise:
  - Classify text into categories
  - Index and search large texts
  - Automatic translation
  - Speech understanding (understand phone conversations)
  - Information extraction (extract useful information from resumes)
  - Automatic summarization (Condense 1 book into 1 page)
  - Question answering
  - Knowledge acquisition
  - Text generation / dialogues

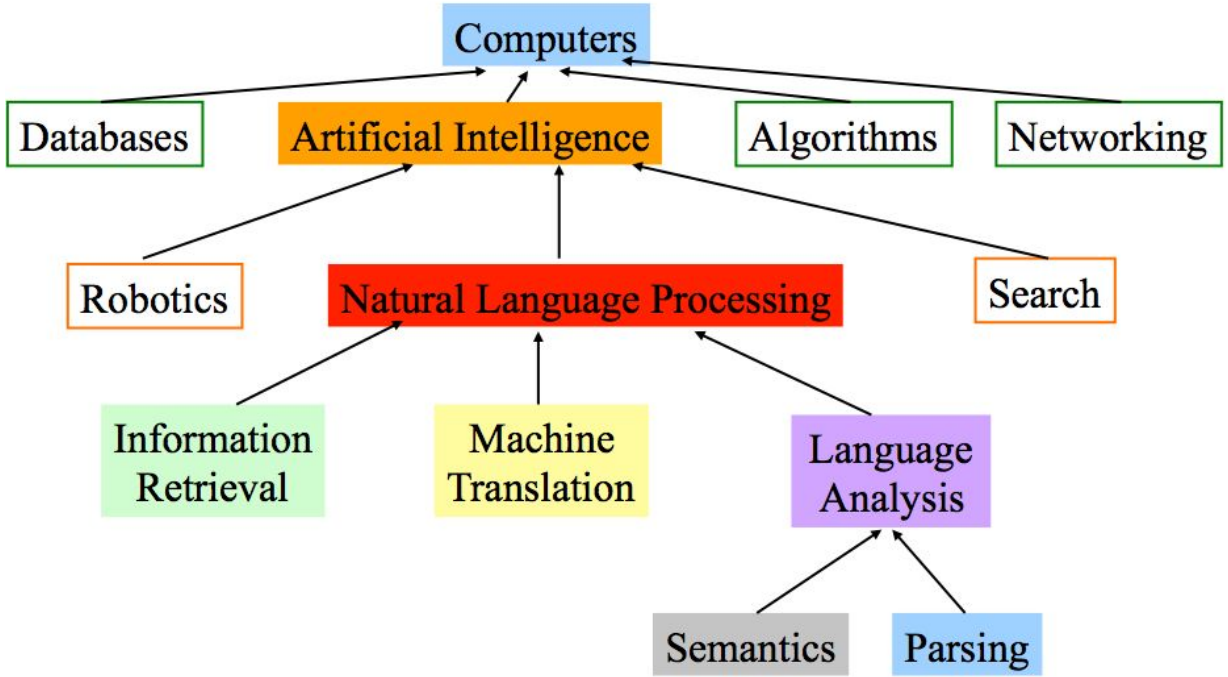
Why Natural Language Processing?

- kJfmmfj mmmvvv nnnffn333
- Uj iheale eleee mnster vensi credur
- Baboi oi cestnitze
- Coovoel2^ ekk; ldsllk lkdf vnnjfj?
- Fgmflmllk mlfm kfre xnnn!

Computers lack knowledge!

- People have no trouble understanding language
  - Common sense knowledge
  - Reasoning capacity
  - Experience
- Computers have
- No common sense knowledge
- No reasoning capacity

Where does NLP fit in the CS taxonomy?



Levels of Analysis

- ★ Speech
  - Phonology
- ★ Text
  - Morphology: the structure of words
  - Syntax: how these sequences are structured
  - Semantics: meaning of the strings
  - Pragmatics: discourse
  - Interaction between levels

Some NLP applications

- ★ Speech recognition
  - Voicemail transcription
- ★ Dialogue systems
  - Siri, Cortana
- ★ Information extraction
  - Named Entity Recognition and Linking
  - Event detection
- ★ Machine translation
  - Text to text
  - Speech to speech

Challenges in NLP

Main issue is ambiguity ….

Ambiguity in Speech

- ★ 264 Lane Street vs. 26 four-lane street
- ★ For invoices vs foreign voices
- ★ Colorectal cancer risks vs co-director cancel risks
- ★ Frapuccino vs Fred Paccino

Ambiguity in Morphological Analysis

ride	rideable
do	doable
like	likeable

- Pattern: Verb + “able” → Adjective (able to do/be Verb-ed)

Ambiguity in Morphological Analysis

happy	unhappy
cool	uncool
stable	unstable

- Pattern: “un” + Adjective → Adjective (not Adjective)

Ambiguity in Morphological Analysis

do	undo
zip	unzip
dress	undress

- Pattern: “un” + Verb → Adjective (to reverse Verb-ing)

Ambiguity in Morphological Analysis

What about the word unlockable?

Image source: http://plywoodchair.com/wp-content/uploads/2015/03/Garage-Door-Lock-Mechanism-Mesmerizing-Door-Design-Ideas.jpg

Ambiguity in Morphological Analysis

Option 1:

“un” + lock (Verb) → unlock (Verb) (to reverse locking)

unlock + “able” → unlockabe (Adjective) (able to unlock)

Option 2:

lock + “able” → lockable (Adjective) (able to lock)

“un” + lockable → unlockable (Adjective) (not able to lock)

Ambiguity in Syntax

Ambiguity in Syntax

- ★ Jake told Mike he has cancer
- ★ Eat spaghetti with meatballs vs eat spaghetti with chopsticks
- ★ We saw the Eiffel Tower flying to Paris
- ★ Old men and women

More Issues in Syntax

Anaphora resolution:

"The dog entered my room. It scared me"

Preposition attachment:

"I saw the man in the park with a telescope"

Issues in Semantics

Understand language, but how?

- ★ “plant” = industrial plant
- ★ “plant” = living organism

Words are ambiguous by design

What is the relevance of getting the semantics right?

- ★ Machine translation (wrong translations)
- ★ Information retrieval (wrong information)

Challenges in Information Extraction

The closure of California's main coastal road demonstrates just how badly these mudslides have damaged this picturesque seaside town, which is simultaneously reeling from the flooding-related deaths of at least 20 residents following the storm early Tuesday morning.

The storm destroyed at least 65 homes and damaged at least 460 more, authorities said. Firefighters are continuing their painstaking work of combing through the debris with heavy equipment and hand tools, aware more bodies are likely buried beneath. At least four people remain missing.

How many deaths? 20? 65? 460? Where? California's coastal road

How many houses destroyed? 20? 65? 460?

Challenges in Information Extraction

Detect new patterns:

- ★ Detect hacking / hidden information / etc./
- ★ Gov. mil, puts lots of money into IE research

Challenges in Information Retrieval

- ★ General model:
  - A huge collection of texts
  - A query
- ★ Task: find documents that are relevant to the given query
- ★ How? Create an index, like the index in a book
- ★ More …
  - Vector-space models
  - Boolean models
- ★ Examples: Google, Yahoo, Baidu, etc.

Challenges in Information Retrieval

- ★ Retrieve Specific Information
- ★ Question Answering:
  - ○ What's the age of the Earth?
  - ○ Approx. ~ 4.5 billion years
- ★ Cross Language Information Retrieval
- ★ What's the minimum age requirement for car rental in Italy?
- ★ Integrate large number of languages
- ★ Integrate into performant IR engines

So Far...

Lots of interesting challenges that require NLP to address them!

What we'll learn this semester

- ★ Some linguistic basics
  - ○ Structure of English
  - ○ Parts of speech, phrases, parsing
- ★ N-grams
  - ○ Also multi-word expressions
- ★ Part of speech tagging
- ★ Syntactic parsing
- ★ Semantics
  - ○ Word sense disambiguation
- ★ Dialogue Systems
- ★ Other higher level NLP tasks

Logistics and Administrivia

Class website:

Piazza website (this is where you go to post questions):
piazza.com/uh/spring2018/cosc6336/home

More Administrivia

Official Book is Jurafsky and Martin 3rd Edition:

http://web.stanford.edu/~jurafsky/slp3/

Please be sure to read the corresponding chapter before class!

Tentative schedule

Week	ADMINISTRIVIA, INTRODUCTION	Reading
1	Linguistics Background & Text Processing, Edit Distance	J&M Ch. 2
2	Language models & Classification	J&M Ch. 4th, 6th & 7th 3rd ed.
3	HMMs and POS tagging	J&M Ch. 9 and 10th 3rd ed.
4	Vector Semantics and word embeddings	J&M Ch. 15th and 16th 3rd ed.
5	Word Senses and HMMs	J&M 18th and 8th 3rd ed.
6	Formal Grammars and Syntactic Parsing	J&M Ch. 12. and 13th

Tentative schedule

Week	ADMINISTRIVIA, INTRODUCTION	Reading
7	Statistical Parsing and Dependency Parsing	J&M Ch. 13th and 14th 3rd. Ed.
8	IE	J&M Ch. 21 3rd ed.
9	QA	J&M Ch. 28th 3rd.
10	Dialog systems	J&M Ch. 29 3rd. Edition
11	Semantic Role Labelling	J&M Ch. 22nd 3rd ed.
12	Other NLP tasks	

Grading Criteria

- ★ 40% - Assignments (mini projects, 3-4 total)
- ★ 40% - Exams (Tentative dates: 1st exam: 3/7/18, second exam: 5/9/18)
- ★ 20% - In-class participation, quizzes and paper presentations

Assignments:

- ★ HW1: Part of speech tagging (write and train your own HMM)
- ★ HW2: Parsing (write and train your own parser)
- ★ HW3: Dialogue system

Exams:

- ★ Practical (Cocalc)

Class requirements

- ★ Bring laptop fully charged to class
- ★ Create a github account and become familiar with it

First Assignment: In class exercise

- From the NLTK book complete the following exercises:
  -