CS 6120/CS4120: Natural Language Processing

Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

Time and Location
- **Time**: Mondays from 6:00 pm – 9:00 pm
- **Location**: Behrakis Health Sciences Cntr 315

Course Webpage

Prerequisites
- Programming
  - Being able to write code in some programming languages (e.g., Python, Java, C/C++, Matlab) proficiently
- Courses
  - Algorithms
  - Some calculus
  - Probability and statistics
  - Linear algebra (optional but highly recommended)

Prerequisites
- A quiz:
  - 22 simple questions, 20 of them as True or False questions (relevant to probability, statistics, and linear algebra)
  - The purpose of this quiz is to indicate the expected background of students.
  - 80% of the questions should be easy to answer.
  - Not counted in your final score!

Textbook and References
- **Main textbook**
  - We will use some material from 3rd edition when it is available.
  - Youtube video: [https://www.youtube.com/watch?v=s3kKUJa3b0](https://www.youtube.com/watch?v=s3kKUJa3b0)
- **Other reference**
  - Chris Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999

- **Machine learning textbooks**:
Topics of the Course

- Language Modeling
- Part-of-Speech Tagging
- Text Categorization: Word Sense Disambiguation, Named Entity Recognition
- Syntax: Formal Grammars of English, Syntactic Parsing, Statistical Parsing, Neural Parsing, Dependency Parsing
- Semantics: Vector Space, Lexical Semantics, Semantics with Nested Structures, Semantic Role Labeling, Semantics Parsing, Combining Logical and Distributional Semantics
- Information Extraction
- Question Answering
- Machine Translation
- Summarization
- Discourse Analysis
- Sentiment Analysis, Opinion Mining
- NLP and Social Media
- Chatbot Systems and Chatsbots

The Goal

- Study fundamental tasks in NLP
- Learn state-of-the-art techniques
- Acquire hands-on skills for solving NLP problems
  - Even some research experience!

Grading

- Assignment
  - 3 assignments, 10% for each
- Quiz
  - 9 in-class tests, 1% for each
- Exam
  - 1 exam, 35%
- Project
  - 1 project, 23%
- Participation
  - 3%
  - Classes, Piazza

Exam

- Open book
- Dec 4, 2017

Course Project

- An NLP-related research project
- 2-3 students as a team

Course Project Grading

- We want to see novel projects!
  - The problem needs to be well-defined, novel, useful, and practical.
  - Reasonable results and observations.
  - We encourage you to tackle a research-driven problem.
- Option 1: a research project discussed with the instructor
  - A better solution for an existing problem
  - On a novel problem
- Option 2: The fake news challenge
Sample Projects from Northeastern NLP Group

Meeting Summarization and Effectiveness Prediction

Questions
• How to pinpoint and extract salient content from a meeting?
• How to understand and model the effectiveness of a meeting?

Understanding and Detecting Supporting Arguments for Contentious Issues
• Alice wants to write an essay
• Topic: Video game and youth violence
• Thesis: Video game contributes to youth violence

Problem
• Alice uses Google to find some related documents.
• But, wait....
Problem

• How to efficiently select relevant arguments?
• Can we ensure the arguments cover different aspects and are non-redundant?

Debate Prediction

“I felt Sanders yelled better on big banks and political corruption but Clinton yelled better on Israel and the minimum wage.”

How Does One Win a Debate?

• Ideally, win a debate based on the merits
  • Facts
  • Reasons
  • Mutual understanding

How Does One Win a Debate?

• However, in reality...

The Joint Effect:
A Discussion on "Abolishing the Death Penalty"

Pro: ... When you look at capital convictions, you can demonstrate on innocence grounds a 4.1 percent error rate... I mean, would you accept that in flying airplanes? ...

Con: ... The risk of an innocent person dying in prison and never getting out is greater if he's sentenced to life in prison than it is if he's sentenced to death. So the death penalty is an important part of our system.

The Joint Effect:
A Discussion on "Abolishing the Death Penalty"

Topic: execution of the innocents

Pro: ... When you look at capital convictions, you can demonstrate on innocence grounds a 4.1 percent error rate... I mean, would you accept that in flying airplanes? ...

Con: ... The risk of an innocent person dying in prison and never getting out is greater if he’s sentenced to life in prison than it is if he’s sentenced to death. So the death penalty is an important part of our system.
The Joint Effect: A Discussion on “Abolishing the Death Penalty”

Pre: ... When you look at capital convictions, you can demonstrate on innocence grounds a 4.3 percent error rate... I mean, would you accept that in flying airplanes...?

Con: ... The risk of an innocent person dying in prison and never getting out is greater if he’s sentenced to life in prison than it is if he’s sentenced to death. So the death penalty is an important part of our system.
Comment Target Extraction

- **Motivation:** companies are interested in how users review or comment on their product or public figures, as a first step, they need to know what users are talking about.
- **Goal:** given a knowledge base of possible entities, and a comment on a news article, or some product review, identify which specific part of the comment refers to which entity.
- **Example:**
  - comment: As a Real Madrid fan, I'm glad that one of the member of the fearsome trident, MSN, will leave LaLiga
  - entity: Real Madrid - [Real Madrid C.F.], MSN - [Lionel Messi, Luis Suárez, Neymar da Silva Santos Júnior], LaLiga - [La Liga Santander]
- **Challenges:**
  - how to detect mention of entity (sometimes one mention can link to multiple entity, such as "they" or "MSN")
  - how to make use of context information (related entities, such as Hillary and Donald)
  - how to jointly do mention detection and entity linking

Option 2: The Fake News Challenge

- **Website:** http://www.fakenewschallenge.org/
- **Goal:** "The goal of the Fake News Challenge is to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem. We believe that these AI technologies hold promise for significantly automating parts of the procedure human fact checkers use today to determine if a story is real or a hoax."

The Fake News Challenge

- **Data:** https://github.com/FakeNewsChallenge/fnc-1

The Fake News Challenge

- **Stage 1: Stance Detection**
  
  **Input:** A headline and a body text - either from the same news article or from two different articles.
  
  **Output:** Classify the stance of the body text relative to the claim made in the headline into one of four categories:
  1. Agrees: The body text agrees with the headline.
  2. Disagrees: The body text disagrees with the headline.
  3. Discusses: The body text discusses the same topic as the headline, but does not take a position.
  4. Unrelated: The body text discusses a different topic than the headline.

Headline: "Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"

```
... Led Zeppelin's Robert Plant turned down $800 MILLION to reunite supergroup ... * 

CORRECT CLASSIFICATION: AGREED
```

```
... No, Robert Plant did not rip up an $800 million Led Zeppelin reunion deal ... *
```

```
CORRECT CLASSIFICATION: DISAGREED
```

```
Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal ... *
```

```
CORRECT CLASSIFICATION: UNRELATED
```

```
Robert Plant's Virgin Galactic is on its launch (Spaceport) Two today ... *
```

CORRECT CLASSIFICATION: DISAGREED
Course Project Grading

- We want to see novel projects!
  - The problem needs to be well-defined, novel, useful, and practical.
  - Reasonable results and observations.
- We encourage you to tackle a research-driven problem.
  - A better solution for an existing problem
  - Or a novel problem
- If you want to do a research project, please talk to the instructor during office hours.

Course Project Grading

- Three reports
  - Proposal (3%)
  - Progress, with code (5%)
  - Final, with code (10%)
- One presentation
  - In class (5%)

Audience Award

- Bonus points!
  - All teams vote for their favorite project(s).
  - Each team votes for their favorite and second favorite
  - Best project gets 2% as bonus, and runner-up gets 1%.

Submission and Late Policy

- Assignment or report turned in late will be charged 20 points (out of 100 points) off for each late day (i.e. 24 hours).
- Each student has a budget of 5 days throughout the semester before a late penalty is applied.
- Late days are not applicable to final presentation.
- Each group member is charged with the same number of late days, if any, for their submission.

Submission and Late Policy

- Each assignment or report, both electronic copy and hard copy, is due at the beginning of class on the corresponding due date.
- Programming language
  - Python, Java, C/C++
- Electronic version
  - On blackboard
  - In class

How to find us?

- Course webpage:
- Office hours
  - Prof. Lu Wang: Mondays from 4:30pm to 5:30pm, or by appointment, 258 WWH
  - TA Tirthraj Maheshkumar Parmar: Tuesdays from 1pm to 2pm, 462 WWH
  - TA Vishruth Krishna Prasad: Wednesdays from 3pm to 4pm, 462 WWH
  - Piazza
  - http://piazza.com/northeastern/fall2017/cs6120/home
  - All course relevant questions should go here – also is the best way to reach the instructor and TAs!
What is Natural Language Processing?
• Allowing machines to communicate with human
• Natural language understanding + natural language generation

What does it mean to understand a language?

Phonology
Morphology
Lexemes
Syntax
Semantics
Pragmatics
Discourse

Sound waves
Words
Parse trees
Meanings

Phonology
Morphology
Lexemes
Syntax
Semantics
Pragmatics
Discourse

Shallower Analysis
Deeper Analysis

Syntax, Semantic, Pragmatics

• Syntax concerns the proper ordering of words and its affect on meaning.
  • The dog bit the boy.
  • The boy bit dog the.
• Semantics concerns the (literal) meaning of words, phrases, and sentences.
  • “plant” as a photosynthetic organism
  • “plant” as a manufacturing facility
  • “plant” as the act of sowing
• Pragmatics concerns the overall communicative and social context and its effect on interpretation.
  • The ham sandwich wants another beer. (co-reference, anaphora)
  • John thinks vanilla. (ellipsis)

Ambiguity is Ubiquitous

• Speech Recognition
  • “recognize speech” vs. “wreck a nice beach”
  • “yesh in Asia” vs. “Kathmandu”
• Syntactic Analysis
  • “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”
• Semantic Analysis
  • “The dog is in the pen” vs. “The ink is in the pen.”
  • “I put the plant in the window” vs. “Ford put the plant in Mexico”
• Pragmatic Analysis
  • From “The Pink Panther Strikes Again”:
    • Clouseau: Does your dog bite?
    • Hotel Clerk: No.
    • Clouseau: [bowing down to pet the dog] Nice doggie.
    • Clouseau: I thought you said your dog did not bite!
    • Hotel Clerk: That is not my dog.
Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in $n$ prepositional phrases has over $2^n$ syntactic interpretations (cf. Catalan numbers).
  - “I saw the man with the telescope”: 2 parses
  - “I saw the man on the hill with the telescope”: 5 parses
  - “I saw the man on the hill in Texas with the telescope”: 14 parses
  - “I saw the man on the hill in Texas with the telescope at noon”: 42 parses
  - “I saw the man on the hill in Texas with the telescope at noon on Monday”: 132 parses

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
  - Policeman to little boy: “We are looking for a thief with a bicycle.” Little boy: “Wouldn’t you be better using your eyes.”
  - Why is the teacher wearing sunglasses? Because the class is so bright.
  - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I’ll never know.
  - She criticized my apartment, so I knocked her flat.
  - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.

Why is Language Ambiguous?

- Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long.
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.
- Language relies on people’s ability to use their knowledge and inference abilities to properly resolve ambiguities.
- Infrequently, disambiguation fails, i.e. the compression is lossy.

Some NLP Tasks

Syntactic Tasks

- Breaking a string of characters into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g., ; : ( )]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar ⇒ myspace . com plucker swing bar
  ⇒ mypace . com pluckers wing bar

Word Segmentation

- Breaking a string of characters into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g., ; : ( )]
Morphological Analysis

- Morphology is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A morpheme is the smallest linguistic unit that has semantic meaning (Wikipedia)
- e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried -> carry + ed (past tense)
  - independently -> in + depend + ent + ly
  - Googlers -> (Google + er) + s (plural)
  - unlockable -> un + lock + able ?
    -> [un + lock] + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.
  - I ate the spaghetti with meatballs.
    Pro  V  Det  N  Prep  N
  - John saw the saw and decided to take it to the table.
    PN  V  Det  N  Con  V  Part  V  Prep  Det  N
- Useful for subsequent syntactic parsing and word sense disambiguation.

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP] [VP and] [NP the spaghetti] [PP with] [NP meatballs],
  - [NP] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only $1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.

Semantic Tasks

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong [interest] in computational linguistics.
  - Ellen pays a large amount of [interest] on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.
Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.
  - agent, patient, source, destination, instrument
- John drove Mary from Austin to Dallas in his Toyota Prius.
- The hammer broke the window.
- Also referred to as “case role analysis,” “thematic analysis,” and “shallow semantic parsing.”

Semantic Parsing

- A semantic parser maps a natural-language sentence to a complete, detailed semantic representation (logical form).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:
  
  How many cities are there in the US?
  answer(A, count(B, (city(B), loc(B, C), const(C, countryid(USA))), A))

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.
- E.g., “A soccer game with multiple males playing.” -> Some men are playing a sport.”

Pragmatics/Discourse Tasks

Anaphora Resolution/Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. Bush did not need the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back.”

More Application-driven Tasks
Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.
- Relation extraction identifies specific relations between entities.
  - Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
  - Who is the president of United States?
    - Donald Trump
  - What is the popular of Massachusetts?
    - 6.8 million

Text Summarization

- Produce a short summary of one or many longer document(s).
  - Article: An international team of scientists studied diet and mortality in 135,335 people between 35 and 70 years old in 18 countries, following them for an average of more than seven years. Diet information depended on self-reports, and the scientists controlled for factors including age, sex, smoking, physical activity and body mass index. The study is in The Lancet. Compared with people who ate the lowest 20 percent of carbohydrates, those who ate the highest 20 percent had a 28 percent increased risk of death. But high carbohydrate intake was not associated with cardiovascular death. ... 
  - Summary: Researchers found that people who ate higher amounts of carbohydrates had a higher risk of dying than those who ate more fats.

Spoken Dialogue Systems -- Chatbots

- Q: Is it going to rain today?
  - A: It will be mostly sunny. No rain is expected.

Machine Translation

- Translate a sentence from one natural language to another.
  - 我喜欢汉堡 → I like burgers.

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - "John plays the guitar." → "John 弹吉他" 
  - "John plays soccer." → "John 踢足球"
  - An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
    - "The spirit is willing but the flesh is weak." → "The liquor is good but the meat is spoiled."
    - "Out of sight, out of mind." → "invisible idiot."
Resolving Ambiguity

• Choosing the correct interpretation of linguistic utterances requires (commonsense) knowledge of:
  • Syntax
    • An agent is typically the subject of the verb
  • Semantics
    • Michael and Ellen are names of people
    • Austin is the name of a city (and of a person)
    • Toyota is a car company and Prius is a brand of car
  • Pragmatics
  • World knowledge
    • Credit cards require users to pay financial interest
    • Agents must be animate and a hammer is not animate

State-of-the-Arts

• Learning from large amounts of text data (cf. rule-based methods)
  • Supervised learning or unsupervised learning
  • Statistical machine learning-based methods
    • The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.
  • Now with neural network-based methods mostly

Related Fields

• Artificial Intelligence
• Machine Learning
• Linguistics
• Cognitive science
• Logic
• Data science
• Political science
• Education
• ...many more

Relevant Scientific Conferences and Journals

• Association for Computational Linguistics (ACL)
• North American Association for Computational Linguistics (NAACL)
• Empirical Methods in Natural Language Processing (EMNLP)
• International Conference on Computational Linguistics (COLING)
• Conference on Computational Natural Language Learning (CoNLL)
• Transactions of the Association for Computational Linguistics (TACL)
• Journal of Computational Linguistics (CL)