# 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**: Tuesdays and Fridays, 3:25 pm - 5:05 pm
- **Location**: West Village H 108

## Course Webpage

- http://www.ccs.neu.edu/home/luwang/courses/cs6120_sp2018/cs6120_sp2018.html

## 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:
  - This Friday, in class
  - 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 (and some slides)
  - Dan Jurafsky and James H. Martin, "Speech and Language Processing, 2nd Edition", Prentice Hall, 2009.
  - We will use some material from 3rd edition when it is available.
    - http://web.stanford.edu/~jurafsky/slp3/
  - Youtube video: https://www.youtube.com/watch?v=s3kKlUBa3b0
- Other reference
  - Chris Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999
- Machine learning textbooks:
  - Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.
  - Tom Mitchell, "Machine Learning", McGraw Hill, 1997.

## Topics of the Course (tentatively)

- Language Modeling
- Part-of-Speech Tagging
- Text Categorization: Word Sense Disambiguation, Named Entity Recognition
- Syntax: Formal Grammars of English, Syntactic Parsing, Statistical Parsing, Dependency Parsing
- Semantics: Vector-Space, Lexical Semantics, Semantics with Dense Vectors
- Information Extraction
- Question Answering
- Machine Translation
- Summarization
- Sentiment Analysis, Opinion Mining
- NLP and Social Media
- Dialog Systems and Chatbots

## The Goal

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

## Grading

- Assignment (30%)
  - 2 assignments, 15% for each
- Quiz (5%)
  - 8 in-class tests, 1% for each (three lowest scores are dropped)
  - Tuesdays, and starting next week
- Final Exam (35%)
- Project (25%)
- Participation (5%)
  - Classes: ask and answer questions, participate in discussions…
  - Piazza: help your peers, address questions…

## Exam

- Open book
- Time and place, TBD (still scheduling with the college)
- Please don't make travel arrangements for exam weeks.

## 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
    - Or a novel problem
  - **Option 2**: The fake news challenge

## Sample Projects from Previous Offering

- Project reports are listed here: http://www.ccs.neu.edu/home/luwang/courses/cs6120_fa2017.html
  - Neural Semantic Parsing Natural Language into SQL
  - Short Passages Reading Comprehension and Question Answering
  - Political Promise Evaluation (PPE)
  - Predicting Personality Traits using Tweets
  - STORY NEXT 2.0: A TEXT INSIGHTS/VISUALIZATION TOOL
  - Android Application for Visual QA
  - Novel Summarizer and Keyword Identifier Using Text Rank with Sentence Farm Detection
  - Paraphrase Generation
  - Hashtag Similarity based on Tweet Text
  - Stance Detection for the Fake News Challenge
  - Machine Comprehension Using match-LSTM and Answer-Pointer
  - Online Abuse Detection
  - Plagiarism Detection Using FP-Growth Algorithm
  - An Examination of Influential Framing of Controversial Topics on Twitter

## Neural Semantic Parsing Natural Language into SQL

- Enterprise stores data in structured format
- Skilled workforce required to extract knowledge
- What if anyone could ask questions in natural language from structured text?

Table: CFLDraft

| Pick # | CFL Team | Player | Position | College |
|---|---|---|---|---|
| 27 | Hamilton Tiger-Cats | Connor Healy | DB | Wilfrid Laurier |
| 28 | Calgary Stampeders | Anthony Forgone | OL | York |
| 29 | Ottawa Renegades | L.P. Ladouceur | DT | California |
| 30 | Toronto Argonauts | Frank Hoffman | DL | York |
| ... | ... | ... | ... | ... |

Question: How many CFL teams are from York College?

SQL: `SELECT COUNT CFL Team FROM CFLDraft WHERE College = "York"`

Result: 2

An example in WikiSQL. The inputs consist of a table and a question. The outputs consist of a ground truth SQL query and the corresponding result from execution.

## Short Passages Reading Comprehension and Question Answering

Answer a simple question by reading a short passage.

- Simple question: Factoid QA, < 30 words.
- Short passage: < 300 words.
- The answer is directly given as a **range** of the passage.

➔ INPUT: A passage, a question.
➔ OUTPUT: a range of the passage.

... John went to school at 9AM and ate a sandwich for lunch and came back home at 5PM. ...

What did John have for lunch?

...In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence...

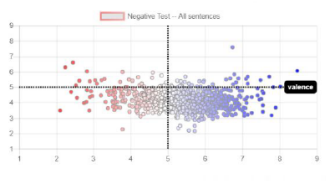
Who published "Computing Machinery and Intelligence"?

## Online Abuse Detection

Automating the process of identifying abuse comments would not only save time for the Social Media platforms but also would increase user safety and improve discussions online.

Build a classifier that classifies the test data as either an "Abuse" (aggression, personal attack or a toxic statement) or "Not-Abuse" statement using multiple techniques.

## StoryNext 2: Sentiment Analysis for Documents



## Option 2: The Fake News Challenge

How Pizzagate went from fake news to a real problem for a D.C. business

By Joshua Gillin on Monday, December 5th, 2016 at 5:23 p.m.


## 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

- 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 discuss the same topic as the headline, but does not take a position
4. **Unrelated**: The body text discusses a different topic than the headline

## The Fake News Challenge

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

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

*"... Led Zeppelin's Robert Plant turned down £500 MILLION to reform supergroup. ..."*

**CORRECT CLASSIFICATION: AGREE**

*"... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ..."*

**CORRECT CLASSIFICATION: DISAGREE**

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

**CORRECT CLASSIFICATION: DISCUSSES**

*"... Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today. ..."*

**CORRECT CLASSIFICATION: UNRELATED**

## 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
  - Feel free to talk to the instructor or TAs on project topics during office hours.

## Course Project Grading

- Three reports
  - Proposal (3%), due by the end of January
  - Progress, with code (7%)
  - Final, with code (10%)
- One presentation
  - In class (5%)

## Audience Award

- Bonus points!
  - All teams vote for their favorite project(s).
  - Best project gets 1% as bonus (one best project in each batch, if we need to have more than one batch/lecture for presentation)

## Submission and Late Policy

- Each assignment or report is due at the beginning of class on the corresponding due date.
- Programming language
  - Python (encouraged), Java, C/C++
- Electronic version
  - On blackboard

## 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.

## How to find us?

- Course webpage:
  - http://www.ccs.neu.edu/home/luwang/courses/cs6120_sp2018/cs6120_sp2018.html
- Office hours
  - Prof. Lu Wang: Tuesdays, from 5:15pm to 6:15pm, or by appointment, 258 WVH
  - TA Liwen Hou
  - TA Tirthraj Maheshkumar Parmar
  - TA Manthan Thakar
- Piazza
  - http://piazza.com/northeastern/sp2018/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?

- "Stop"
- "Turn it up"
- "Volume level 6"
- "Repeat that"
- "What can you do?"
- "Play some music"
- "Play music by [artist]"
- "Play dance music on YouTube"
- "Play KEXP radio on TuneIn"
- "Play the latest episode of Radiolab"
- "Pause"
- "Next song"
- "When's my first appointment tomorrow?"
- "Wake me up at 6am tomorrow"
- "Tell me about my day"
- "How long will it take to get to work?"
- "What's the weather today?"

## What does it mean to understand a language?

Phonology
Morphology
Lexemes
Syntax
Semantics
Pragmatics
Discourse

Sound waves
↓
Words
↓
Parse trees
↓
Meanings

## What does it mean to understand a language?

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 the dog.
  - Bit boy dog the 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.
  - John thinks vanilla.

[Modified from Ray Mooney's Slides]

## Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"

## Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- Syntactic Analysis
  - "I ate spaghetti with chopsticks" vs. "I ate spaghetti with meatballs."

## Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- 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"

## Ambiguity is Ubiquitous

- Speech Recognition
  - "recognize speech" vs. "wreck a nice beach"
  - "youth in Asia" vs. "euthanasia"
- 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.
    [*Dog barks and bites Clouseau in the hand*]
    **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 sun-glasses. 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?

## 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

## 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. ,;.-:()]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar
    ⇒ myspace .com pluckers wing bar
    ⇒ myspace .com plucker swing bar

## 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 Pro Prep Det N

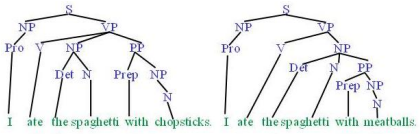
- Useful for subsequent syntactic parsing and word sense disambiguation.

## Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He ] [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 a "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. But the president needed 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.

  people organizations places
  - 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.
  - 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"

## 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
    - Some social norm, communicative goals
    - Asking a question, expecting an answer
  - 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)