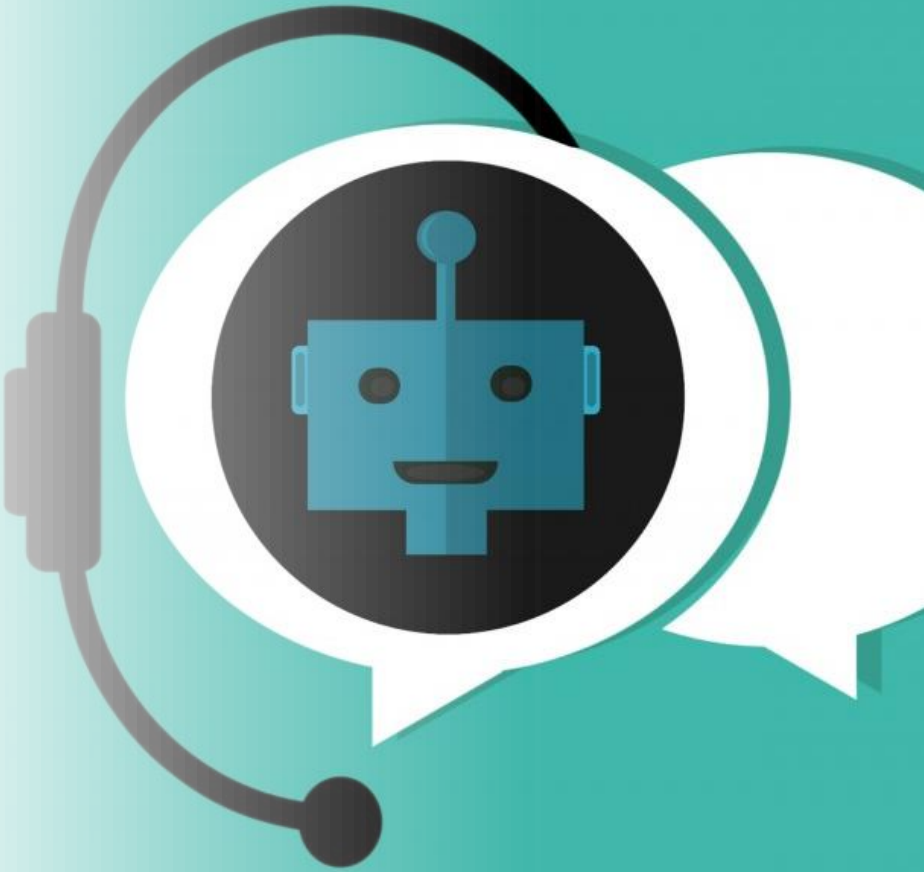
Lecture 1 – Introduction to Natural Language Processing

Jackie Chi Kit Cheung

Fall 2023 - COMP 550

Readings: J&M Chapter 1



Who Am I?

Associate Professor at McGill	2021 -
• Associate Scientific Co-Director at Mila	
Assistant Professor at McGill	2015 – 2021
PhD in Computer Science (Toronto)	2014

Research topics in my lab

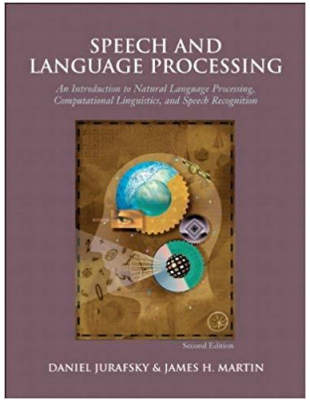
- Natural language generation
- Automatic summarization
- Computational semantics
- Computational pragmatics
- Applications of NLP

Preliminaries

Instructor:	Jackie Chi Kit Cheung
Time and Loc.:	TR 13:05 – 14:25, ENGMC 304
Office hours:	T 14:30 – 16:30, ENGMC 108N
TAs:	Ali Saheb Pasand
	Elaine Lau
	Mikael Brunila
	Hao Yuan Bai
Evaluation:	2 programming assignments (20%)
	4 reading assignments (20%)
	1 midterm (20%)
	1 group project (40%)

Textbook

Jurafsky and Martin. *Speech and Language Processing* (2nd edition)



E-book available online through myCourses
Hard copy available at bookstore
Draft chapters of 3rd edition available online:
https://web.stanford.edu/~jurafsky/slp3/

Assignments

Two programming assignments (10% each x 2 = 20%)

- Hand in online
- Programming to be done in Python 3.

Four reading assignments (5% each x 4 = 20%)

- Covers advanced material and applications

Midterm

Worth 20% of your final grade

To be completed in person.

Awaiting central scheduling services.

Aiming for **week of November 7**

More details as we approach the midterm date.

Final Project

Worth 40%.

Experiment on some language data set

Summarize and review relevant papers

Report on experiments

Must be done in teams of three

Coming up with a project idea:

- Extend a model we see in class
- Work on a relevant topic of interest
- Consult a list of suggested projects, to be posted

Project Steps

Paper or project proposal

Progress update

Final submission

Due dates to be announced

General Policies

Lateness policy for assignments:

- Grace period of 24 hours
- > 24 hours: accepted if it is convenient for us at our discretion

Plagiarism: just don't do it—I regularly catch and submit cases.

Language policy: In accord with McGill policy, you have the right to write essays and examinations in English or in French.

Generative AI Usage

Fine to use in an assistive manner

- Help understand course content
- Search for information
- Brainstorm ideas
- Edit writing

Must acknowledge use of this technology.

Not okay to use as primary means to complete tasks

- Feed in assignment questions to generate solutions
- Generate project report from scratch on a topic

Platforms

Ed Discussions

You'll be added soon

Most releases will be done via this platform

myCourses

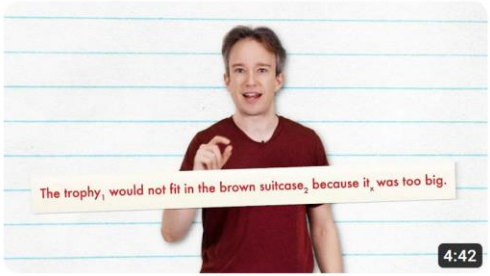
Assignment and project submissions

Grade release

Computational Linguistics and Natural Language Processing

Large Language Models – Impressive Impact!

- Question answering, code generation, essay writing, summarization
- Commercial uses: customer service, personal assistants, healthcare
- Many informal uses: entertainment, settling disputes



The Sentences Computers Can't Understand, But Humans Can

5M views • 3 years ago

Tom Scott

(Those are affiliate links that give a commission to me or Gretchen, depending on country!) REFERENCES: Levesque, H.J., Davis, ...

CC

Tom Scott, 2020

"Artificial language processing remains 10 years away, just as it has for the last few decades."

Tom Scott, 2023

"… that this new technology, the thing that was going to change everything, was starting to actually change everything"

I tried using AI. It scared me.

5.1M views • 13 days ago

Tom Scott

Script assistant: Laura Conlon No AI assistance was used, except where ...

4K CC

Intro | I just wanted to fix my email |... 6 chapters

How Do Language Models Work?

Key insight: learn correlations between words in context

Language modelling:

Mary had a little _____

- *lamb* GOOD
- *accident* GOOD?
- *very* BAD
- *up* BAD

Do this at internet-scale with sophisticated statistical techniques (deep learning)!

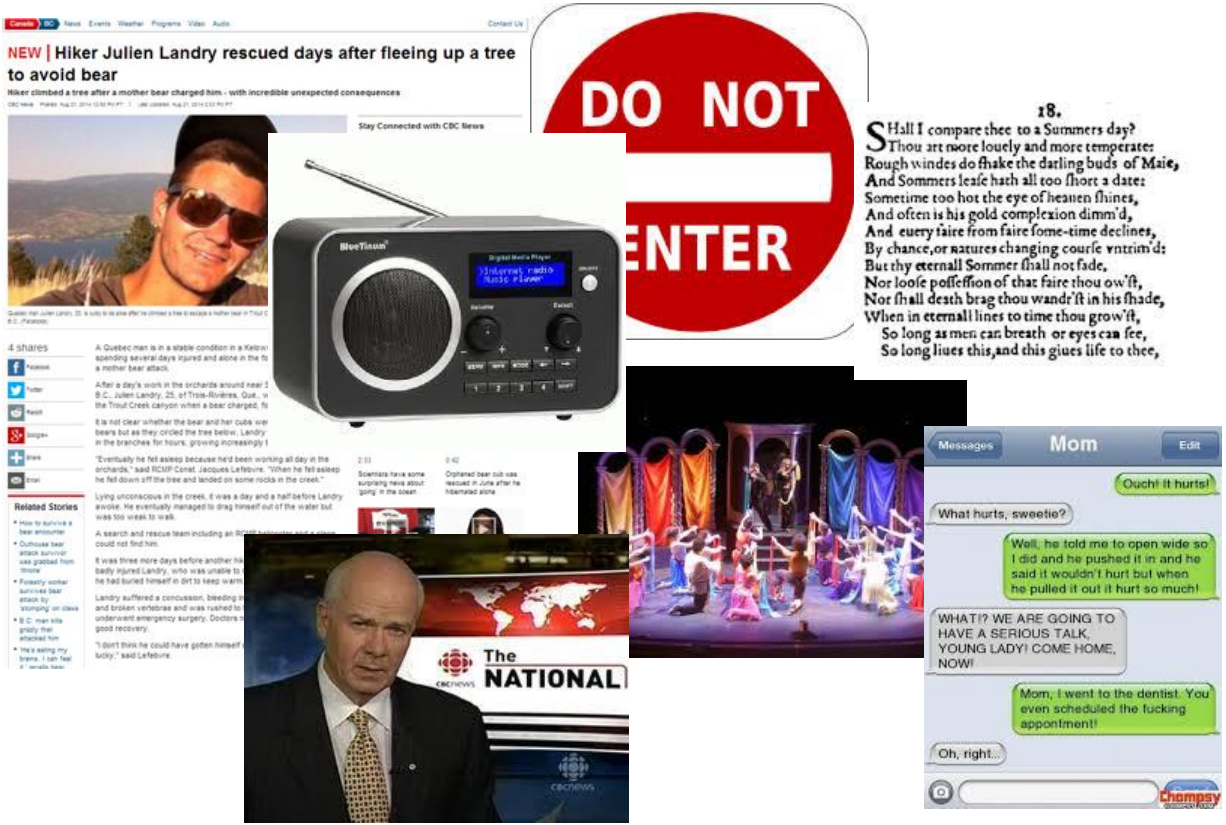
What This Course Is About

- How did we get to large language models dominating NLP research?
- What was the progression of the field of NLP? Why did people try the methods that they did?
- What are some common tasks and paradigms involving natural language?
- How do we evaluate and analyze NLP systems?
- How are properties of natural language reflected in NLP research?

What This Course Is Not About

- The latest techniques in language modelling
- Deep learning / machine learning as a primary focus
  - We will touch on this, and you can do a final project that uses ML, but it is **not** the primary focus of the course.

What Is Language Anyway?



Languages Are Diverse

6000+ languages in the world

language

langue

ਭਾਸ਼ਾ

語言

idioma

Sprache

lingua

→lingyourlanguage

https://lingyourlanguage.com/ (My high score is 513 on Omniglot)

What is Language?

Some properties:

- Form of communication
- **Arbitrary** pairing between form and meaning
- Primarily vocal (exception: sign languages)
- Highly expressive and productive
- Nearly universal (barring developmental disorders)

How do these compare?

- Programming language (e.g., C, Python, Java)
- Vocalizations by your favourite animal
- Written English

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Domains of natural language

- Acoustic signals, phonemes, words, syntax, semantics, …
- Speech vs. text
- **Natural language understanding** (or **comprehension**) vs. **natural language generation** (or **production**)

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Goals

- Language technology applications
- Scientific understanding of how language works

Computational Linguistics (CL)

Modelling natural language with computational models and techniques

Methodology and techniques

- Gathering data: language resources
- Evaluation
- Statistical methods and machine learning
- Rule-based methods

Natural Language Processing

Computational linguistics and **natural language processing (NLP)** are sometimes used interchangeably.

Slight difference in emphasis:

NLP	**CL**
Goal: practical technologies	Goal: how language actually works
Engineering	Science

Understanding and Generation

Natural language understanding (NLU)

Language to form usable by machines or humans

- E.g., parsing, sentiment analysis

Natural language generation (NLG)

Traditionally, semantic formalism to text

More recently, also text to text

- E.g., machine translation, chatbots

Personal Assistant App

Understanding

Call a taxi to take me to the airport in 30 minutes.

What is the weather forecast for tomorrow?

Generation

Machine Translation

I like natural language processing.

↕

Automatische Sprachverarbeitung gefällt mir.

Understanding

Generation

Computational Linguistics

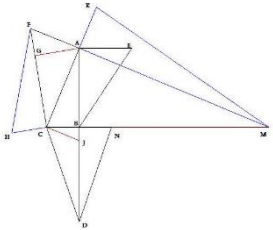
Besides new language technologies, there are other reasons to study CL and NLP as well.

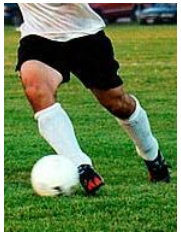
The Nature of Language

First language acquisition

Chomsky proposed a **universal grammar**

Is language an “instinct”?





What innate knowledge must children already have in order to learn their mother tongue, given their exposure to linguistic inputs?

Train a model to find out!

The Nature of Language

Language processing

Some sentences are supposed to be grammatically correct, but are difficult to process.

Formal mathematical models to account for this.

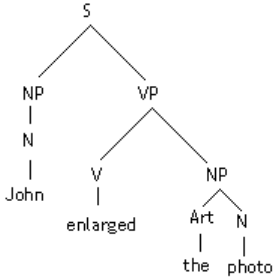
The rat escaped.

The rat the cat caught escaped.

?? *The rat the cat **the dog chased** caught escaped.*

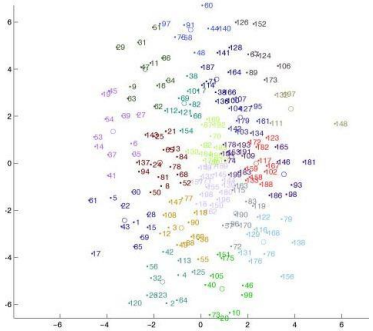
Mathematical Foundations of CL

We describe language with various formal systems.



cat + z > cats

cat + z	*SS	Agree	Max	Dep	Ident
catiz				*!	
catis				*!	*
catz		*!			
cat			*!		
☞ cats					*



$$\begin{bmatrix} \text{phon } \textit{made} \\ \text{synsem} \mid \text{loc} \mid \text{cat} \begin{bmatrix} \text{head} \begin{bmatrix} \text{maj } v \\ \text{vform } \textit{finite} \end{bmatrix} \\ \text{valence} \begin{bmatrix} \text{subj} \langle \text{NP} \rangle \\ \text{compls} \langle \text{VP}[\textit{base}], \text{NP} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} \text{phon } \textit{force} \\ \text{synsem} \mid \text{loc} \mid \text{cat} \begin{bmatrix} \text{head} \begin{bmatrix} \text{maj } v \\ \text{vform finite} \end{bmatrix} \\ \text{valence} \begin{bmatrix} \text{subj} \langle \text{NP} \rangle \\ \text{compls} \langle \text{VP}[\textit{inf}], \text{NP} \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Mathematical Foundations of CL

- Mathematical properties of formal systems and algorithms
  - Can they be efficiently learned from data?
  - Efficiently recovered from a sentence?
  - Complexity analysis
- Implications for algorithm design

Domains of Language

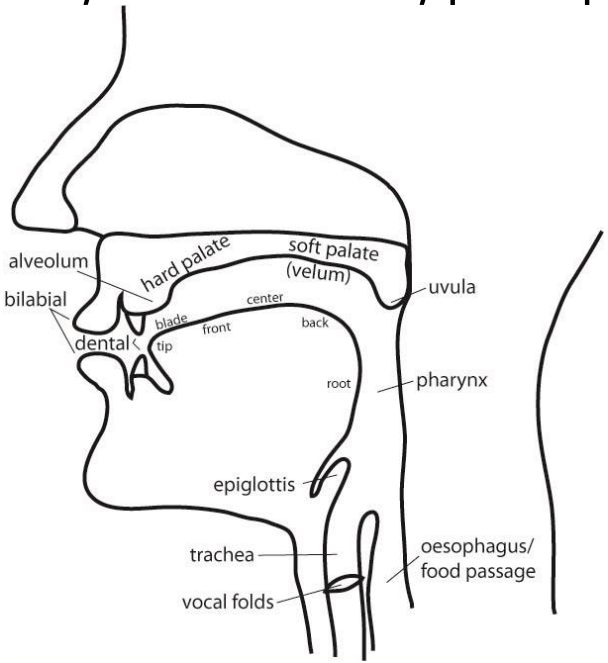
The grammar of a language has traditionally been divided into multiple levels.

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Phonetics

Study of the speech sounds that make up language

Articulation, transmission, perception



peach

[phi:tsh]

Involves closing of the lips, building up of pressure in the oral cavity, release with aspiration, …

Vowel can be described by its formants, …

Phonology

Study of the rules that govern sound patterns and how they are organized

peach	[phi:tsh]	/pi:t͡ʃ/
speech	[spi:tsh]	/spi:t͡ʃ/
beach	[bi:tsh]	/bi:t͡ʃ/

The p in peach and speech are the same phoneme, but they actually are phonetically distinct!

Morphology

Word formation and meaning

antidisestablishmentarianism

anti- dis- establish -ment -arian -ism

establish

*establish**ment***

*establishment**arian***

*establishmentarian**ism***

***dis**establishmentarianism*

***anti**disestablishmentarianism*

Syntax

Study of the structure of language

**I a woman saw park in the.*

I saw a woman in the park.

The first sentence is not well formed (it is **ungrammatical**), while the second one is.

- Words must be arranged in a certain order in a certain way to be a valid English sentence!

Syntax

http://explosm.net/comics/1682/

There are two meanings for the first sentence in the comic! What are they? This is called **ambiguity**.

Semantics

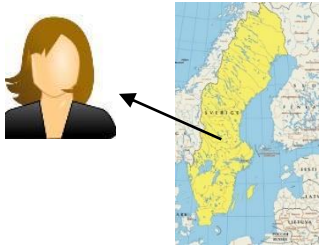
Study of the meaning of language

bank

Ambiguity in the **sense** of the word

Semantics

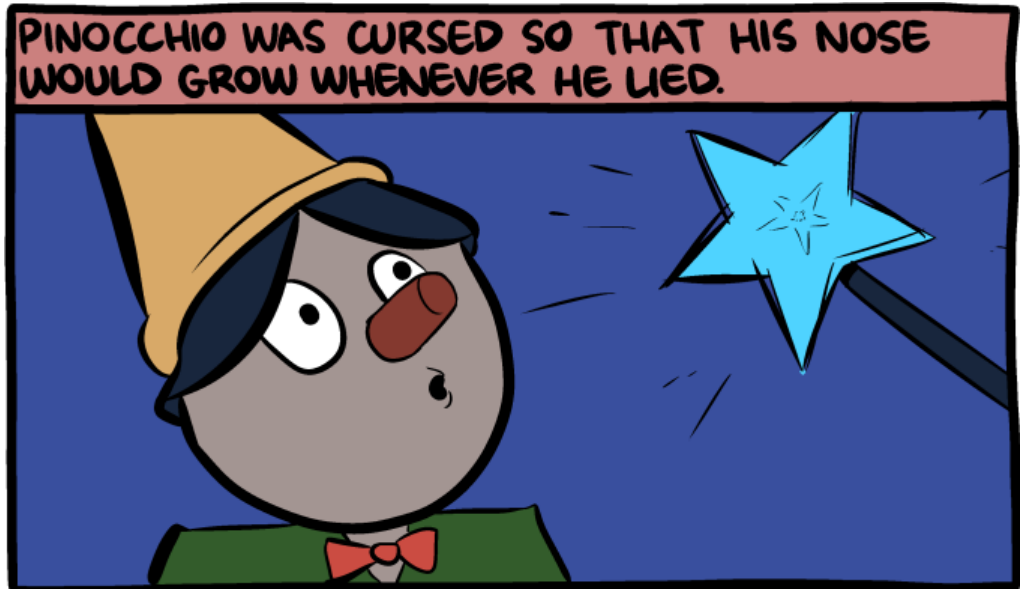
Ross wants to marry a Swedish woman.



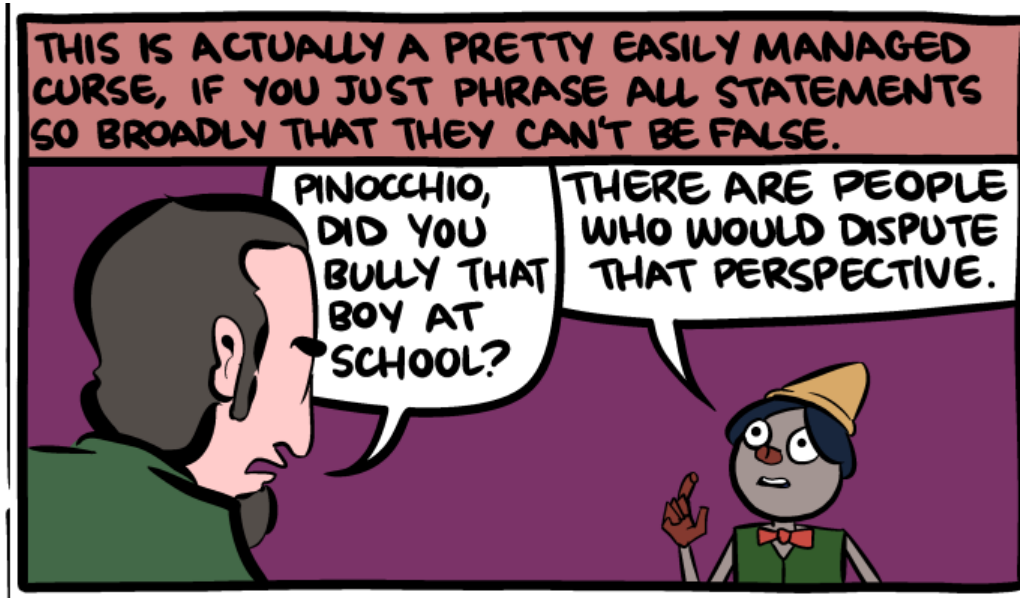
Pragmatics

Study of the meaning of language in context.

→ Literal meaning (semantics) vs. meaning in context:
http://www.smbc-comics.com/index.php?id=3730



Pragmatics


THIS IS ACTUALLY A PRETTY EASILY MANAGED CURSE, IF YOU JUST PHRASE ALL STATEMENTS SO BROADLY THAT THEY CAN'T BE FALSE.
PINOCCHIO, DID YOU BULLY THAT BOY AT SCHOOL?
THERE ARE PEOPLE WHO WOULD DISPUTE THAT PERSPECTIVE.

Pragmatics


YOU CAN ALSO ANSWER TOUGH QUESTIONS BY JUST DESCRIBING THE CURRENT SITUATION.
PINOCCHIO! DID YOU EGG MY DOOR?
IF YOU'RE ACCUSING ME OF SUCH A DEED, I HAVE NOTHING MORE TO SAY.

Pragmatics


CURSED TO NEVER LIE, PINOCCHIO CEASED TO INTERACT WITH TRUTH, EVEN IN SAFE SITUATIONS.
SIR, WOULD YOU LIKE ANY DESSERT?
DESSERT WOULD BE DELICIOUS.

Pragmatics – Deixis

Interpretation of expressions can depend on **extralinguistic** context

e.g., pronouns

I think cilantro tastes great!

The entity referred to (the **antecedent**) by *I* depends on who is saying this sentence.

Shameless Plug – COMP 767 in Winter 2024

Consider taking my topics course in Winter 2024:

Formal and Neural Models of Pragmatics

Read and discuss classic and modern papers on computational pragmatics

- Evaluation based on paper presentations and research project

Discourse

Study of the structure of larger spans of language (i.e., beyond individual clauses or sentences)

I am angry at her.
She lost my cell phone.

I am angry at her.
The rabbit jumped and ate two carrots.

NLP – the Technological Perspective

A combination of **pre-specified knowledge** and **machine learning from data**

Problem specification
Machine learning algorithms
Human annotations
Linguistic knowledge
…

Websites
News articles
Discussions
Knowledge bases
…

NLP Tools and Techniques

Major paradigms for NLP, not mutually exclusive:

Rule-based systems

- Often hand-engineered knowledge about language
- E.g., *heureux* -> *happy*

Machine learning

- Model learns about language through examples
- **Classification**: e.g., is this e-mail spam?
- **Sequence models**: make series of decisions
- Many other paradigms

Knowledge representation

- Formal structure to encode what model knows
- Logic? A large set of continuous-valued numbers?

Topics in COMP-550

Organized roughly by level of linguistic analysis and a corresponding technical approach (ML or otherwise)

NLP Topic	Linguistic layer	Techniques
Text classification	Words	Classification
Language modelling, POS tagging	Words (esp. syntactic structure of words)	Sequence models
Syntactic parsing	Syntactic structure	Structure prediction, dynamic programming
Computational semantics, coreference resolution	Meaning (semantics, discourse)	Logic, semi-supervised learning, neural models
Applications: MT, summarization, etc.	Various	Various

Applications in COMP-550

Last three weeks of the course focus on language technology applications and advanced topics:

- Vision and language
- Automatic summarization
- Machine translation
- Evaluation issues in NLP

Accompanied by reading assignments!

Feel free to send me suggestions of topics you would like to see covered.

Course Objectives

- Understand the broad topics, applications and common terminology in the field
- Prepare you for research or employment in CL/NLP
  - Learn some basic linguistics
  - Learn the basic algorithms
  - Be able to read an NLP paper
- Understand the challenges in CL/NLP
  - Answer questions like "Is it easy or hard to…"
  - Evaluate claims made by companies, the press, and others about CL/NLP