Introduction to Natural Language Processing

COMP-599

Sept 2, 2016

Preliminaries

Instructor: Jackie Chi Kit Cheung

Time and Loc.:   MW 13:05-14:25 in MC 103

Office hours: W 14:30-15:30 or by appointment in MC108N

TA: Jad Kabbara

Evaluation: 4 assignments (40%)

                   1 midterm (20%)

                   1 project (30%)

                   1 oral final exam (10%)

The Course Is Full

There are currently at least 8 people trying to get into the course.

If you're leaning towards dropping the course, please do it soon, so that others can get in! (Though we'll miss you!)

Due to resource limits, I cannot extend the class size anymore.

Alternative: LING 550

- LING 550 is offered by Morgan Sonderegger. There may be one or two spots for *graduate students*.

LING 550 vs COMP 599

LING 550 does not assume background in programming or probability theory; DOES assume background in basic linguistics

LING 550 covers more of:

- Computational methods for linguistic analysis
- Speech; automatic speech recognition

COMP 599 covers more of:

- Algorithms, machine learning models for NLP
- Technological applications of NLP

General Policies

Lateness policy for assignments: no late assignments accepted.

Plagiarism: just don't do it.

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

Course website:
http://cs.mcgill.ca/~jcheung/teaching/fall-2016/comp599/index.html

Important announcements given in-class or on course website, not on MyCourses

Assignments

Four assignments (10% each)

Involve readings, problem sets and programming component.

  Programming component – hand in online through myCourses

  Programming to be done in Python 2.7.

  Non-programming components – hand in on paper in class

Midterm

Worth 20% of your final grade

Currently scheduled for Wed, November 9, 2016

Will be conducted in-class (80 minutes long). More details as we approach the midterm date.

Final Project

Worth 30%.

Experiment on some language data set

Summarize and review a few relevant papers

Report on experiments

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

Final Examination (10%)

Presentation of your final project

7 minute presentation + 3 minute QA session

Four sessions, to be scheduled during final exam period, which you will sign up for

Computational Linguistics and Natural Language Processing

Language is Everywhere

12

Languages Are Diverse

6000+ languages in the world

language

langue

ਭਾਸ਼ਾ

語言

idioma

Sprache

lingua

→The Great Language Game

http://greatlanguagegame.com/ (My high score is 1300)

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

Sometimes, **computational linguistics** and **natural language processing (NLP)** are 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

Natural language generation (NLG)

    Traditionally, semantic formalism to text

    More recently, also text to text

Most work in NLP is in NLU

    c.f. linguistics, where most theories deal primarily with production

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

Automatic Summarization

We want to condense the information in some source text or texts.

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"?
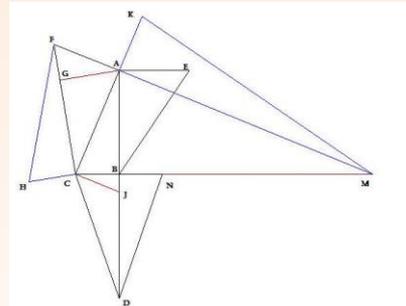
Do children have enough linguistic input to learn their mother tongue?

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 <u>the cat caught</u> escaped.

?? *The rat the cat <u>**the dog chased** caught</u> escaped.*

Mathematical Foundations of CL

We describe language with various formal systems.
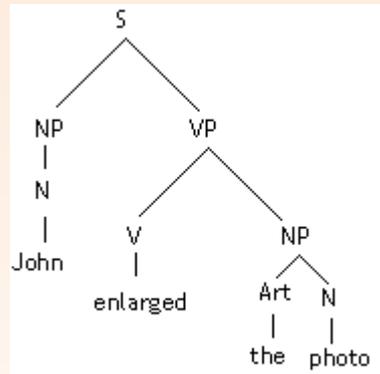
cat + z > cats					
cat + z	*SS	Agree	Max	Dep	Ident
catiz				*!	
catis				*!	*
catz		*!			
cat			*!		
☞ cats					*

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

$$\left[\begin{array}{l} \text{phon } force \\ \text{synsem} \mid \text{loc} \mid \text{cat} \left[\begin{array}{l} \text{head} \left[\begin{array}{l} \text{maj } v \\ \text{vform } finite \end{array}\right] \\ \text{valence} \left[\begin{array}{l} \text{subj} \langle \text{NP} \rangle \\ \text{compls} \langle \text{VP}[inf], \text{NP} \rangle \end{array}\right] \end{array}\right] \end{array}\right]$$
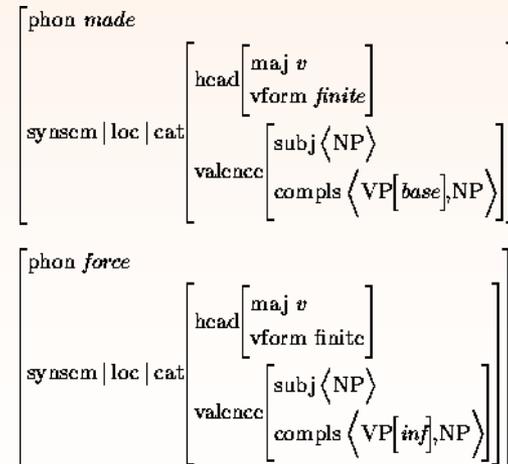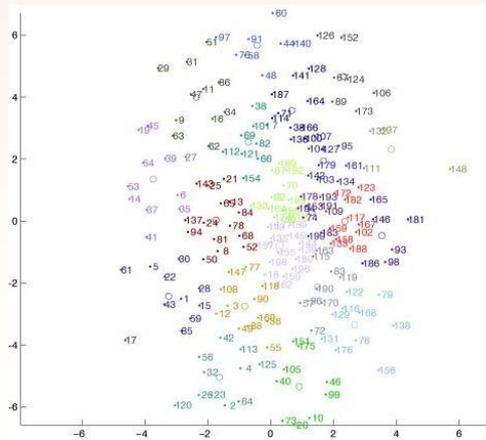
Mathematical Foundations of CL

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

Types of Language

Text

Much of traditional NLP work has been on news text.

Clean, formal, standard English, but very limited!

More recent work on diversifying into multiple domains

Political texts, text messages, Twitter

Speech

Messier: disfluencies, non-standard language

Automatic speech recognition (ASR)

Text-to-speech generation

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
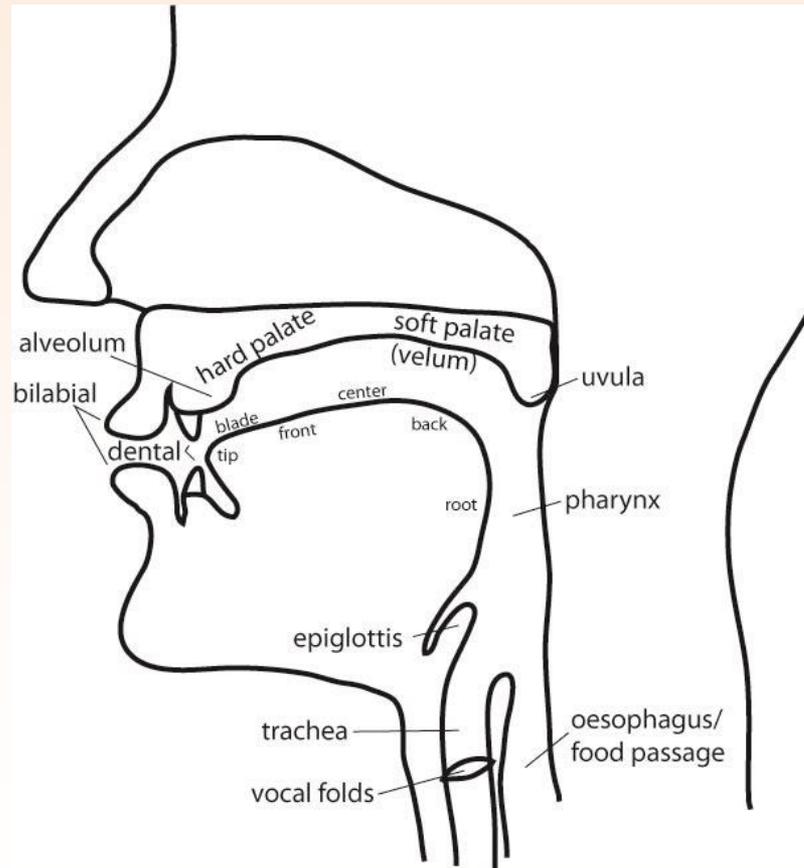
peach   [phi:tsh]

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]

speech   [spi:tsh]

beach   [bi:tsh]

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**

disestablishmentarianism

antidisestablishmentarianism

Syntax

Study of the structure of language

>*I a woman saw park in the.

>I saw a woman in the park.

There are two meanings for the sentence above! 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 <u>a</u> 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
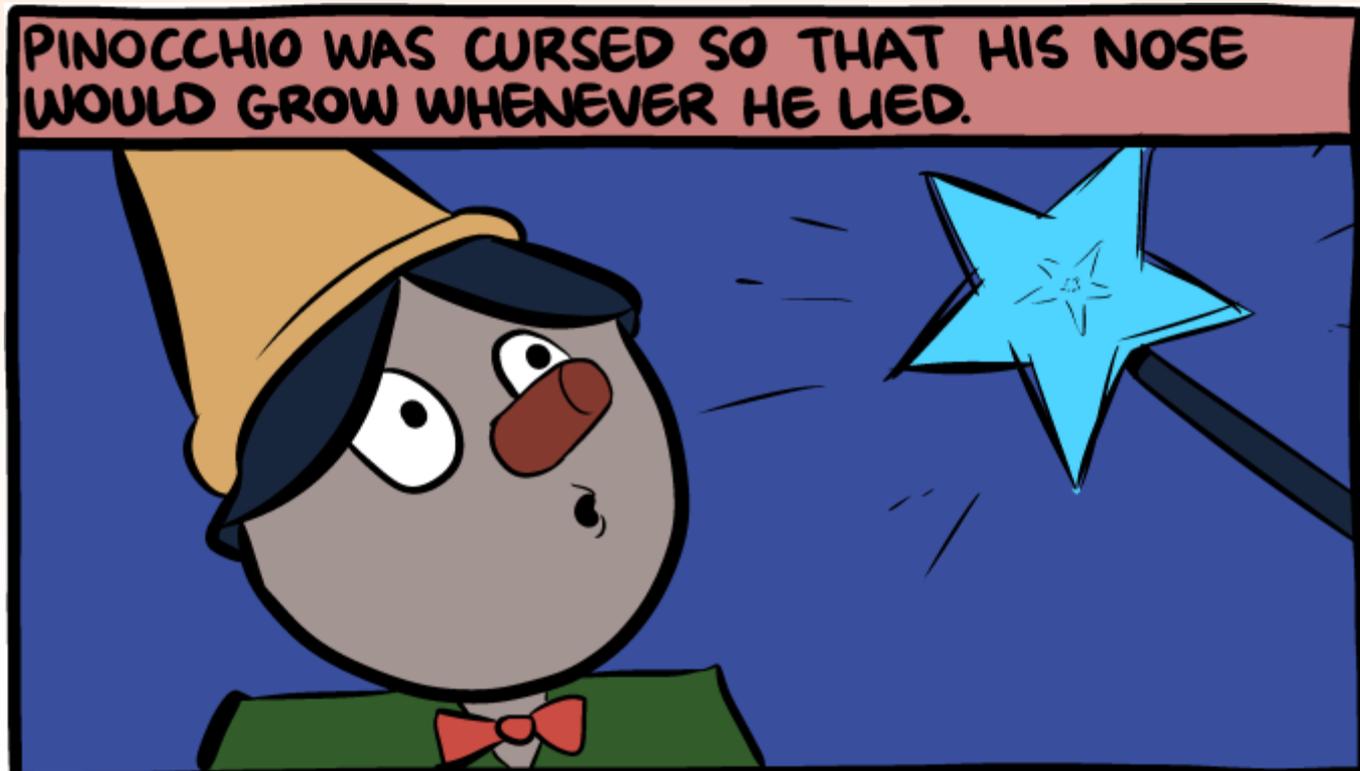
Pragmatics

Pragmatics

Pragmatics

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.

Questions

1. What is the difference between phonetics and phonology?

2. What are two possible readings of this phrase? What level does the ambiguity act at? (i.e., lexical, syntactic, semantic, discourse)

   - *old men and women*

A Brief History of Computational Linguistics

Beginnings in Machine Translation

Early researchers in the 1950s were wildly optimistic.

Georgetown-IBM experiment:

A demonstration of Russian to English MT, featuring 6 translation rules and knowledge of around 250 words in the two languages.

This resulted in substantial interest and funding for MT

Researchers thought that with a little bit more work in engineering the rules and a more complete dictionary of words, they could develop a passable system. They were wrong.

→ http://www.hutchinsweb.me.uk/AMTA-2004.pdf

Disillusionment and the AI Winter

The **Automatic Language Processing Advisory Committee (ALPAC)** report came out in 1966.

- Criticized MT research and its future prospects

- Its effect was to reduce funding to MT and NLP in general, which continued into the seventies.

- The current name for the Association for Computational Linguistics was changed from the Association for Machine Translation and Computational Linguistics in 1968.

Part of the AI winter, in which funding and interest in AI research stagnated

If You're Interested

PopSci YouTube videos on the topic:

https://www.youtube.com/watch?v=5sLbWItc33I

https://www.youtube.com/watch?v=2ac41CO7Nr0

Handcrafted Rule-based Systems

Up until the late 1980s, much work in CL involved coming up with formal analyses of natural language using carefully designed rules.

This led to very precise systems that could give you lots of information about the small fragment of language it knows about, but which are limited in domain and scope.

The Statistical Revolution

Starting in the late 80s, early 90s, the trend became to learn grammar rules from data, rather than specify them.

Often, the level of analysis was shallower, so that it would be something that could be learned by simple statistical models.

Algorithms developed to get the analysis with the highest probability according to some statistical model. Use this to resolve ambiguity.

Machine learning and **empirical evaluation** on **corpora** of naturally occurring language samples became very important.

Modern Trends

Continuation of statistical revolution

More sophisticated machine learning techniques

Make better use of the large amounts of language data available

Require less supervision or input from humans to learn useful regularities in language.

New applications for the Internet age

Real-time language translation

Semantic search to directly access information

Sentiment analysis to predict trends

<Your brilliant idea here>

Main Organizations and Venues

Association for Computational Linguistics

ACL, NAACL, EACL, EMNLP (Empirical Methods in Natural Language Processing), CoNLL (Conference on Natural Language Learning)

Workshops of associated special interest groups

All publications are open-access on the ACL Anthology! http://aclweb.org/anthology/

Others:

COLING, IJCNLP ("Asian ACL")

Journals

Computational Linguistics, Natural Language Engineering, ACM/IEEE Transactions on Audio Speech and Language Processing

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 to…"; see through hype

This Semester in COMP-599

We'll progress through the subfields, roughly organized by the level of linguistic analysis

Morphology -> Syntax -> Semantics -> Discourse

We'll cover selected NLP applications in more details in the last part of the course.

Along the way:

Learn some basic linguistics

Learn algorithms to analyze linguistic structure

Learn some machine learning techniques for the above

Reminder

No class on Monday, Labour Day!