CMSC 473/673
Natural Language Processing
Fall 2018
Frank Ferraro

ITE 358
ferraro@umbc.edu
Monday: 2:15-3
Tuesday: 11:00-11:30
by appointment

Natural language processing
Semantics
Vision & language processing
Learning with low-to-no supervision
Caroline Kery

Location TBD
ckery1@umbc.edu
Tuesday: 2-3:30pm
Thursday: 1-2:30pm
by appointment

Semantic parsing
Active learning
Data visualization
Analysis of educational data
The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

BY SIDON LEWIS-KRAUS  DEC. 14, 2016
August 2018

Artificial Intelligence Is Now a Pentagon Priority. Will Silicon Valley Help?

The Defense Department, believing that A.I. research should be a national priority, has called on the White House to “inspire a whole of country effort.”

By CADE METZ

Aug. 26, 2018

Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?

Digital assistants from Amazon, Apple and Google can only have meager back-and-forth exchanges with us. Listen to how that tells us something about where they’re going in the future.

By KEITH COLLINS and CADE METZ

Aug. 17, 2018

Google Employees Protest Secret Work on Censored Search Engine for China

About 1,400 of the internet company’s employees have signed a letter demanding transparency, saying censored search results raise “urgent moral and ethical issues.”

By KATE CONGER and DAISUKE WAKABAYASHI

Aug. 16, 2018
Potential Applications

ASR (automatic speech recognition)

Machine translation

Natural language generation

Document labeling/classification

Document summarization

Corpus exploration

Relation/information extraction

Entity identification
Automatic speech recognition
Natural language generation

Hi all,
We wanted to invite you to join us for an early Thanksgiving on November 22nd, beginning around 2PM. Please bring your favorite dish! RSVP by next week.

Dave

Hi team,

The server appears to be dropping about 10% of requests (see attached dashboards). There hasn’t been a new release since last night, so I’m not sure what’s going on. Is anyone looking into this?

...
A massive climate change study is canceled because of climate change

By Doug Criss, CNN

Updated 10:37 AM ET, Tue June 20, 2017

Story highlights

Arctic sea ice has traveled farther south than normal along Newfoundland’s northeast coast

An icebreaker has been repeatedly diverted to take part in rescue operations

(CNN) — A $17 million study of climate change in the Canadian Arctic has been nixed for now -- because of climate change.

A team of scientists from the University of Manitoba and four other schools were in the middle of the first leg of a four-year study of how climate change is affecting the areas around the Hudson Bay, the university said in statement. The study, named BaySys, started last month, and the scientists were traveling on the Canadian Research Icebreaker CCGS Amundsen.

But because of warmer temperatures in the Arctic, hazardous sea ice is traveling farther south than usual. The Amundsen, which is part of the Canadian Coast Guard fleet, has been diverted several times because its ice-breaking capabilities have been needed to help out in rescue efforts along Newfoundland’s northeast coast. All of the delays and concerns about safety forced...
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan.

He said Pat would try the same tactic again.
Pat and Chandler agreed on a plan. He said Pat would try the same tactic again.

is “he” the same person as “Chandler?”

Entity identification
Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)
The New York Times

Building the Next New York Times Recommendation Engine

by Alexander Spanger
August 11, 2015 11:27 AM

The Washington Post

Innovations

When artificial intelligence makes a picture worth way more than a thousand words

By Dominic Basulto
December 1, 2015

IBT

Facebook AI Still Can’t Do Things Even A Baby Has Mastered

By Charles Poladian
On 06/16/16 at 5:09 PM

The Huffington Post

Shifting toward the Knowledge Economy

Engadget

Google's AI can translate language pairs it has never seen
Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)

Learn different ways that success and progress can be measured in NLP
Deep Learning
Natural Language Processing

What society thinks I do
What my friends think I do
What other computer scientists think I do

What mathematicians think I do
What I think I do
What I actually do
Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)
Learn different ways that success and progress can be measured in NLP
Relate to statistics, machine learning, and linguistics
Implement NLP programs
Course Goals

Be introduced to some of the core problems and solutions of NLP (big picture)
Learn different ways that success and progress can be measured in NLP
Relate to statistics, machine learning, and linguistics
Implement NLP programs
Read and analyze research papers
Practice your (written) communication skills
T-REX IN: "COMPUTATIONAL LINGUISTICS"
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing!
Natural Language Processing
≈
Computational Linguistics
Natural Language Processing
≈
Computational Linguistics

\textit{science focus}

computational bio
computational chemistry
computational X
Natural Language Processing

≈

Computational Linguistics

engineering focus

science focus

computational bio
computational chemistry
computational X
Natural Language Processing ≈ Computational Linguistics

Machine learning
Natural Language Processing ≈ Computational Linguistics

Machine learning

Information Theory
Natural Language Processing ≈ Computational Linguistics

Machine learning

Information Theory

Data Science
Natural Language Processing ≈ Computational Linguistics

Machine learning

Information Theory

Data Science

Systems Engineering
Natural Language Processing ≈ Computational Linguistics

Machine learning

Information Theory

Data Science

Systems Engineering

Logic

Theory of Computation
Natural Language Processing ≈ Computational Linguistics

- Machine learning
- Information Theory
- Data Science
- Systems Engineering
- Logic
- Theory of Computation
Natural Language Processing ≈ Computational Linguistics

Machine learning

Information Theory

Data Science

Systems Engineering

Logic

Theory of Computation

Linguistics

Cognitive Science

Psychology
Natural Language Processing ≈ Computational Linguistics

Machine learning  Linguistics
Information Theory  Cognitive Science
Data Science  Psychology
Systems Engineering  Political Science
Logic  Digital Humanities
Theory of Computation  Education
Natural Language Processing ≈ Computational Linguistics

build a system to translate
create a QA system

engineering focus

these views can co-exist peacefully
What Are Words?

Linguists don’t agree

(Human) Language-dependent

White-space separation is a sometimes okay (for written English longform)

Social media? Spoken vs. written? Other languages?
What Are Words? Tokens vs. Types

The film got a great opening and the film went on to become a hit.

**Type**: an element of the vocabulary.

**Token**: an instance of that type in running text.

How many of each?
Terminology: Tokens vs. Types

The film got a great opening and the film went on to become a hit.

<table>
<thead>
<tr>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>The</td>
</tr>
<tr>
<td>film</td>
<td>film</td>
</tr>
<tr>
<td>got</td>
<td>got</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>great</td>
<td>great</td>
</tr>
<tr>
<td>opening</td>
<td>opening</td>
</tr>
<tr>
<td>and</td>
<td>and</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>went</td>
<td>went</td>
</tr>
<tr>
<td>on</td>
<td>on</td>
</tr>
<tr>
<td>to</td>
<td>to</td>
</tr>
<tr>
<td>become</td>
<td>become</td>
</tr>
<tr>
<td>hit</td>
<td>hit</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
The film got a great opening and the film went on to become a hit.

**Types**
- The
- film
- got
- a
- great
- opening
- and
- the
- went
- on
- to
- become
- hit
- .

**Tokens**
- The
- film
- got
- a
- great
- opening
- and
- the
- film
- went
- on
- to
- become
- a
- hit
- .
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing.

A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language – English, for example – can have more than one meaning!

Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?
orthography

Adapted from Jason Eisner, Noah Smith
orthography

morphology

Adapted from Jason Eisner, Noah Smith
orthography

morphology

lexemes

Adapted from Jason Eisner, Noah Smith
orthography
morphology
lexemes
syntax
semantics

Adapted from Jason Eisner, Noah Smith
orthography
morphology
lexemes
syntax
semantics
pragmatics
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse

Adapted from Jason Eisner, Noah Smith
NLP + Latent Modeling

explain what you see/annotate

with things “of importance” you don’t

observed text

orthography

morphology

lexemes

syntax

semantics

pragmatics

discourse

Adapted from Jason Eisner, Noah Smith
orthography
morphology
lexemes
syntax
semantics
pragmatics
discourse

VISION
AUDIO

prosody
intonation
color
Language is Productive
Watergate
Troopergate

Watergate ➔ Bridgegate

Deflategate
Language is Ambiguous
Kids Make Nutritious Snacks
Ambiguity

Kids Make Nutritious Snacks
Kids *Prepare* Nutritious Snacks
Kids *Are* Nutritious Snacks

sense ambiguity
British Left Waffles on Falkland Islands
British Left Waffles on Falkland Islands

Ambiguity

lexical ambiguity
British Left Waffles on Falkland Islands

 lexical ambiguity
Parts of Speech

Classes of words that behave like one another in “similar” contexts

Pronunciation (stress) can differ: object (noun: OB-ject) vs. object (verb: ob-JECT)

It can help improve the inputs to other systems (text-to-speech, syntactic parsing)
Pat saw Chris with the telescope on the hill.

I ate the meal with friends.
Pat saw Chris with the telescope on the hill.

I ate the meal with friends.

syntactic ambiguity
Language Can Be Surprising
The old
Garden Path Sentences

The old man
The old man the
The old man the boat
The old man the boat.
The old man the boat.
The complex houses married and single soldiers and their families.
The complex houses married and single soldiers and their families.
The rat the cat the dog chased killed ate the malt.
The rat *that* the cat the dog chased killed ate the malt.
Garden Path Sentences

The rat *that* the cat *that* the dog chased killed ate the malt.
The rat *that* the cat *that* the dog *chased* killed ate the malt.
The rat *that* the cat *that* the dog chased *killed* ate the malt.
The rat *that* the cat *that* the dog chased *killed* ate the malt.
Garden Path
Sentences

[The rat [the cat [the dog chased] killed] ate the malt].

Language can have recursive patterns

**Syntactic parsing** can help identify those
Syntactic parsing: perform a “meaningful” structural analysis according to grammatical rules.

I ate the meal with friends.
Syntactic Parsing Can Help Disambiguate

I ate the meal with friends.
Syntactic Parsing Can Help Disambiguate

I ate the meal with friends.
Clearly Show Ambiguity...
But Not Necessarily All Ambiguity

I ate the meal with a fork
I ate the meal with gusto
I ate the meal with friends
Discourse Processing

John stopped at the donut store.
Discourse Processing

John stopped at the donut store.
John stopped at the donut store before work.
Discourse Processing

John stopped at the donut store on his way home.
Discourse Processing

John stopped at the donut shop.

John stopped at the trucker shop.

John stopped at the mom & pop shop.

John stopped at the red shop.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
Discourse Processing through Coreference

I spread the cloth on the table to protect it.

I spread the cloth on the table to display it.
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing!

A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language – English, for example – can have more than one meaning!

Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?

Many have focused on statistical modelling of language, but this approach is approximate. I agree!
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today.
probabilistic model

\[ p_{\theta}(X) \]

objective

\[ F(\theta) \]
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$
Gradient Ascent

\[ \arg \max_{\theta} F(\theta) \]
Gradient Ascent

$$\arg \max_{\theta} F(\theta)$$

"gradient of F with respect to $\theta$"
"gradient of $F$ with respect to $\theta$"

gradient: a vector of derivatives, each with respect to $\theta_k$ while holding all other variables constant
T-REX IN: "COMPUTATIONAL LINGUISTICS"

Computational linguistics is the study of computer-based language processing!

A major area of computational linguistics is that of "ambiguity resolution". It turns out that many things people say in a language - English, for example - can have more than one meaning!

Consider the phrase "fruit flies like a banana". Is it describing the taste of fruit flies, or rather flying fruit? How can a computer hope to figure this out?

Many have focused on statistical modelling of language, but this approach is approximate. I agree!

What do YOU know about computational linguistics? Ever read a little paper called "Non-Statistical Models for Unsupervised Prepositional Phrase Attachment?"

That was me! It was some of my earliest work on head word tuples!

(C) 2003 Ryan North
http://universaldependencies.org/
part-of-speech & syntax for > 120 languages
From Syntax to Shallow Semantics

“Open Information Extraction”

Angeli et al. (2015)

http://corenlp.run/ (constituency & dependency)

https://github.com/hltcoe/predpatt

http://openie.allenai.org/ (constituency trees)

http://www.cs.rochester.edu/research/knext/browse/ (constituency trees)

http://rtw.ml.cmu.edu/rtw/
Administrivia
## Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>473</th>
<th>673</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five Assignments</td>
<td>45%</td>
<td>30%</td>
</tr>
<tr>
<td>Midterm</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Graduate Paper</strong></td>
<td>---</td>
<td>30%</td>
</tr>
<tr>
<td>Course Project</td>
<td>45%</td>
<td>30%</td>
</tr>
</tbody>
</table>
## Final Grades

<table>
<thead>
<tr>
<th>≥</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>A</td>
</tr>
<tr>
<td>80</td>
<td>B</td>
</tr>
<tr>
<td>70</td>
<td>C</td>
</tr>
<tr>
<td>65</td>
<td>D</td>
</tr>
<tr>
<td>0</td>
<td>F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>≥</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>A-</td>
</tr>
<tr>
<td>80</td>
<td>B-</td>
</tr>
<tr>
<td>70</td>
<td>C-</td>
</tr>
<tr>
<td>65</td>
<td>D</td>
</tr>
<tr>
<td>0</td>
<td>F</td>
</tr>
</tbody>
</table>
Natural Language Processing

Fall 2018 — CMSC 473/673

ANNOUNCEMENTS

- 8/28/17 [473/673] The syllabus is available.

WHO, WHAT, WHEN, AND WHERE

Check out the syllabus (to be released soon) for all this information, including policies on academic honesty, accommodations, and late assignments.

Meeting Times
Sherman Hall, 015
Monday & Wednesday, 1pm - 2:15pm

Instructor
Frank Ferraro
ferraro [at] umbc [dot] edu
ITE 358
Monday 2:15 - 3pm
Tuesday 11:00 - 11:30
by appointment

TA
Caroline Kery
ckery1 [at] umbc [dot] edu
TBD

Topics
The topics covered will include

- probability, classification, and the efficacy of simple counting methods
- language modeling (n-gram models, smoothing heuristics, maximum probability/maximum entropy, and vector valued representations)
- sequences of latent variables (e.g., hidden Markov models, some alignment)
- trees and graphs, as applied to syntax and semantics
- some discourse-related applications (coreference resolution, text summarization)
- special and current topics (e.g., fairness and ethics in NLP)

Goals
After taking this course, you will

- be introduced to some of the core problems and solutions of NLP;
- learn different ways that success and progress can be measured;
- be exposed to how these problems relate to those in statistics, machine learning, and other fields;
- have experience implementing a number of NLP programs;
- read and analyze research papers;
- practice your (written) communication skills.
Online Discussions

https://piazza.com/umbc/fall2018/cmsc473673
## Important Dates

<table>
<thead>
<tr>
<th>Date</th>
<th>473</th>
<th>673</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, 9/10</td>
<td>Assignment 1</td>
<td>Assignment 1</td>
</tr>
<tr>
<td>Monday, 10/1</td>
<td>Assignment 2</td>
<td>Assignment 2</td>
</tr>
<tr>
<td>Friday, 10/12</td>
<td>Project Proposal</td>
<td>Project Proposal</td>
</tr>
<tr>
<td>Wednesday, 10/17</td>
<td>—</td>
<td>Graduate Paper</td>
</tr>
<tr>
<td>Monday, 10/22</td>
<td>Assignment 3</td>
<td>Assignment 3</td>
</tr>
<tr>
<td>Wednesday, 10/31</td>
<td>In-class Midterm</td>
<td>In-class Midterm</td>
</tr>
<tr>
<td>Friday, 11/9</td>
<td>Project Update</td>
<td>Project Update</td>
</tr>
<tr>
<td>Wednesday, 11/14</td>
<td>—</td>
<td>Graduate Paper Reviews</td>
</tr>
<tr>
<td>Monday, 11/19</td>
<td>Assignment 4</td>
<td>Assignment 4</td>
</tr>
<tr>
<td>Wednesday, 12/5</td>
<td>—</td>
<td>Revised Graduate Paper</td>
</tr>
<tr>
<td>Monday, 12/10</td>
<td>Assignment 5</td>
<td>Assignment 5</td>
</tr>
<tr>
<td>Wednesday, 12/19</td>
<td>Final Project</td>
<td>Final Project</td>
</tr>
</tbody>
</table>
Late Policy

Everyone has a budget of 10 *late days*

If you have them left: assignments turned in after the deadline will be graded and recorded, no questions asked

If you don’t have any left: still turn assignments in. They could count in your favor in borderline cases
Late Policy

Everyone has a budget of 10 late days

Use them as needed throughout the course
They’re meant for personal reasons and emergencies

Do not procrastinate
Late Policy

Everyone has a budget of 10 late days

Contact me privately if an extended absence will occur