NATURAL LANGUAGE COMPUTING
What is natural language computing?

Getting computers to understand everything we say and write.

In this class (and in the field generally), we are interested in the statistics of language.

(Occasionally, computer models give insight into how humans process language.)
Today

• Common challenges with natural language processing (NLP).

• Applications
  • Translating between languages
  • Speech recognition
  • Answering questions
  • Summarizing long documents

• My own research.

• Course logistics.
What can natural language do?

The ultimate in **human-computer interaction**.

“translate *Also Sprach Zarathustra*”

“open the pod bay doors”

“how far until Jupiter?”

“Can you summarize *2001: A Space Odyssey*?”

We’re making progress, but why are these things *still* hard to do?
A little deeper

• Language has **hidden structures**, e.g.,
  • How are **sounds** and **text** related?
    • e.g., why is this: not a ‘ghoti’ (enough, women, nation)?

• How are words **combined** to make sentences?
  • e.g., what makes ‘**colourless green ideas sleep furiously**’ **correct** in a way unlike ‘furiously sleep ideas green colourless’?

• How are words and phrases used to produce **meaning**?
  • e.g., if someone asks ‘**do you know what time it is?**’, why is it **inappropriate** to answer ‘**yes**’?

• We need to organize the way we think about language…
Categories of linguistic knowledge

• **Phonology:** the study of patterns of speech sounds.
  e.g., “read” → /r iy d/

• **Morphology:** how words can be changed by inflection or derivation.
  e.g., “read”, “reads”, “reader”, “reading”, ...

• **Syntax:** the ordering and structure between words and phrases (i.e., grammar).
  e.g., \( \text{NounPhrase} \rightarrow \text{article adjective noun} \)

• **Semantics:** the study of how meaning is created by words and phrases.
  e.g., “book” →

• **Pragmatics:** the study of meaning in contexts.
Ambiguity – Phonological

- **Phonology**: the study of patterns of speech sounds.

- "read" → /r iy d/ as in ‘I like to read’
- "read" → /r eh d/ as in ‘She read a book’
- "object" → /aa₁ b jh eh⁰ k t / as in ‘That is an object’
- "object" → /ah⁰ b jh eh¹ k t / as in ‘I object!’
- “too” ← /t uw/ as in ‘too much’
- “two” ← /t uw/ as in ‘two beers’

- Ambiguities can often be **resolved** in context, but not always.
  - e.g., /h aw t uw r eh¹ k ah ?? n ay² z s (b|p) iy ch/ → ‘how to recognize speech’
  → ‘how to wreck a nice beach’
Resolution with syntax

• If you hear the sequence of speech sounds
  \(/b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow\ ...
  b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow\ b\ ah\ fae\ l\ ow/\n
  which word sequence is being spoken?
  → “Buff a low buff a lobe a fellow Buff a low buff a lobe a fellow...”
  → “Buffalo buff aloe buff aloe buff aloe buff aloe buff aloe buff aloe ...”
  → “Buff aloe buff all owe Buffalo buffalo buff a lobe ...”
  → “Buff aloe buff all owe Buffalo buff aloe buff a lobe ...”
  → “Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo”

• It’s obvious (to us) that the last option is most likely because
  we have knowledge of **syntax**, i.e., grammar.
Ambiguity – Syntactic

- **Syntax**: the ordering and structure between words. Words can be grouped into ‘parse tree’ structures given grammatical ‘rules’.

  e.g., “I shot an elephant in my pyjamas”
Resolution with semantics

- It’s obvious (to us) that the elephants don’t wear pyjamas, and we can discount one option because of our knowledge of **semantics**, i.e., meaning.
Ambiguity – Semantic

- **Semantics**: the study of how meaning is created by the use of words and phrases.

- “Every man loves a woman”
  \[ \forall x \text{man}(x) \exists y: (\text{woman}(y) \land \text{loves}(x, y)) \]
  \[ \exists y: \text{woman}(y) \land \forall x (\text{man}(x) \rightarrow \text{loves}(x, y)) \]

- “I made her duck”
  \[ \rightarrow \text{I cooked waterfowl meat for her to eat.} \]
  \[ \rightarrow \text{I cooked waterfowl that belonged to her.} \]
  \[ \rightarrow \text{I carved the wooden duck that she owns.} \]
  \[ \rightarrow \text{I caused her to quickly lower her head.} \]

- “Give me the pot”
  \[ \rightarrow \text{It’s time to bake.} \]
  \[ \rightarrow \text{It’s time to get baked.} \]
Resolution with pragmatics

• It’s obvious (to us) which meaning is intended given knowledge of the context of the conversation or the world in which it takes place.

  • “Every man loves a woman”
    \[\forall x \text{ man}(x) \exists y: (\text{woman}(y) \land \text{loves}(x, y))\]
    \[\exists y: \text{woman}(y) \land \forall x (\text{man}(x) \rightarrow \text{loves}(x, y))\]

  • “I made her duck”
    \[\rightarrow \text{I cooked waterfowl meat for her to eat.}\]
    \[\rightarrow \text{I cooked waterfowl that belonged to her.}\]
    \[\rightarrow \text{I carved the wooden duck that she owns.}\]
    \[\rightarrow \text{I caused her to quickly lower her head.}\]

  • “Give me the pot”
    \[\rightarrow \text{It’s time to bake.}\]
    \[\rightarrow \text{It’s time to get baked.}\]
Ambiguity – miscellaneous

- Newspaper headlines (spurious or otherwise)

  - Kicking Baby Considered to be Healthy
  - Squad Helps Dog Bite Victim
  - Canadian Pushes Bottle Up Germans
  - Milk Drinkers are Turning to Powder
  - Grandmother of Eight Makes Hole in One
  - Kids Make Nutritious Snacks
  - Juvenile Court Tries Shooting Defendant
  - Local High School Dropouts Cut in Half

CSC401/2511 – Spring 2016
NLC as Artificial Intelligence

• NLC involves **resolving ambiguity** at all levels.
  • **Reasoning with world knowledge.**
    • In the early days knowledge was **explicitly encoded** in artificial symbolic systems (e.g., context-free grammars) by **experts**.

• Now, algorithms learn using **probabilities** to distinguish subtly different competing hypotheses.
  • E.g., is *Google* a noun or a verb?
  • An example where *Google* ∈ Nouns ("*Google makes Android*"), does not mean that Google is never a verb ("*Go Google yourself*").

• \( P(\text{Google} \in \text{Nouns}) > P(\text{Google} \in \text{Verbs}) > 0 \)
The Turing Test

• **First** and most fundamental test of machine intelligence.

• A machine (A) imitates a human using nothing but a text-based instant messenger.
  • If a human interrogator (C) cannot reliably differentiate a real human (B) from the machine, that machine is said to be ‘intelligent’.

Aside – Chatterbots

• ELIZA (Weizenbaum, 1966): simple pattern matching to imitate a psychiatrist.
• Surprisingly effective despite no linguistic knowledge.
• e.g.,
  
  User: Men are all alike.
  ELIZA: In what way?
  User: They’re always bugging us about something or other.
  ELIZA: Can you think of a specific example?
  User: My boyfriend made me come here.
  ELIZA: Your boyfriend made you come here.  
  (Jurafsky and Martin, 2009)
Course outline (approximate)

• Introduction and linguistic data (2 lectures)
• N-gram models and features of data (2 lectures) *
• Entropy and information theory (2 lectures) *
• Hidden Markov models (3 lectures) *
• Statistical machine translation (3 lectures) **
• Articulatory and acoustic phonetics (2 lectures) *
• Automatic speech recognition (2 lectures) **
• Speech synthesis (1 lecture) **
• Information retrieval (2 lectures) **
• Text summarization (1 lecture) **
• Other classifiers and review (2 lectures)

* techniques  ** applications
Preview: Machine translation

- One of the most prized applications in NLC.
- Requires both interpretation and generation.
- Over $100B spent annually on human translation.
According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959B US dollars of foreign capital, including 40.007B US dollars of direct investment from foreign businessmen.

IBM4

The Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007B US dollars today provide data include that year to November China actually using foreign 46.959B US dollars and

Yamada/Knight

Today’s available data of the Ministry of Foreign Trade and Economic Cooperation shows that China’s actual utilization of November this year will include 40.007B US dollars for the foreign direct investment among 46.959B US dollars in foreign capital.
Preview: Machine translation

- In the 1950s and 1960s direct word-for-word replacement was popular.
- Due to semantic and syntactic ambiguities and differences in source languages, results were mixed.

"The spirit is willing, but the flesh is weak"

"The vodka is good, but the meat is rotten"
One problem is disparity of meanings in languages.

**nation n.** a large body of people, associated with a particular **territory**, that is sufficiently conscious of its **unity** to seek or to possess a **government** of its own.

**nation n.** an aggregation of persons of the same **ethnic family**, often speaking the same **language** or cognate **languages**.
Preview: Machine translation

• **Solution**: automatically learn statistics on parallel texts

... citizen of Canada has the right to vote in an election of members of the House of Commons or of a legislative assembly and to be qualified for membership ...

e.g., the *Canadian Hansards*: bilingual Parliamentary proceedings

... citoyen canadien a le droit de vote et est éligible aux élections législatives fédérales ou provinciales ...

Statistical machine translation

• Modern statistical machine translation is based on the following perspective...

When I look at an article in Russian, I say: ‘This is really written in English, but it has been **coded** in some strange symbols. I will now proceed to **decode**.’

Warren Weaver  March, 1947

Claude Shannon  July, 1948
Aside – Machine translation

- [http://www.translationparty.com](http://www.translationparty.com) uses Google Translate to go back and forth between English and Japanese until we get two consecutive identical English phrases.
Preview: Machine translation

Start with an English phrase:

that's one small step for a man, one giant leap for mankind

find equilibrium

Step One small step for mankind, this hotel is ideal for men

Equilibrium found!
Okay, I get it, you like Translation Party.
Preview: Speech recognition

- Buy ticket... AC490... yes
- Put this there.
- My hands are in the air.

Dictation

Telephony

Multimodal interaction
Speech waveforms

“Two plus seven is less than ten”
Spectrograms

- Speech sounds can be thought of as overlapping sine waves.
- Speech is split apart into a 3D graph called a ‘spectrogram’.
- Spectrograms allow machines to extract statistical features that differentiate between different kinds of sounds.
Speech recognition

beet /bɪt/  bat /bæt/  bott /bat/  boot /but/
In order to classify an unknown observation (e.g., \( X \)), we need a \textit{statistical} model of the distribution of sounds.
Preview: Information retrieval

Google

WolframAlpha

which woman has won more than 1 nobel prize?

About 503,000 results (0.31 seconds)

Who has won more than one Nobel prize? - The Times of India

1 Apr 2007 ... Who has won more than one Nobel prize? Marie Curie won the Nobel prize in 1903 for Physics and 1911 in Chemistry, Linus Pauling in 1954 (for ... timesofindia.indiatimes.com; Sunday TOI; Open Space; Cached; Similar

Answers.com - Who won the Nobel Prize more than once

Who first won the nobel prize in the world? How many families won the nobel prize? What invention has won a nobel prize? Has someone won more than one nobel ... wiki.answers.com/.../Who_won_the_Nobel_Prize_more_than_once - Cached; Similar

Nobel Prize - Wikipedia, the free encyclopedia

A shoulder and head picture of a woman in her fifties sitting in a chair. ... Charles K. Kao won the 2009 Nobel Prize in Physics for his research in optical fibres. ... There has been one laureate, William Vickers, who died after the prize was ... en.wikipedia.org/wiki/Nobel_Prize - Cached; Similar

List of Nobel laureates - Wikipedia, the free encyclopedia

Six laureates have received more than one prize; of the six ... en.wikipedia.org/wiki/List_of_Nobel_laureates - Cached; Similar

Has anyone ever won Nobel prizes in more than one category over ... wiki.answers.yahoo.com/Arts & Humanities; History - Cached; Similar

Results:

<table>
<thead>
<tr>
<th>year</th>
<th>recipient</th>
<th>field</th>
<th>country of achievement</th>
<th>country of birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>Akira Suzuki</td>
<td>chemistry</td>
<td>Japan</td>
<td>Japan</td>
</tr>
<tr>
<td>2010</td>
<td>Ei-ichi Negishi</td>
<td>chemistry</td>
<td>United States</td>
<td>Japan</td>
</tr>
<tr>
<td>2010</td>
<td>Richard F. Heck</td>
<td>chemistry</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Christopher A. Pissarides</td>
<td>economics</td>
<td>United Kingdom</td>
<td>Cyprus</td>
</tr>
<tr>
<td>2010</td>
<td>Dale T. Mortensen</td>
<td>economics</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Peter A. Diamond</td>
<td>economics</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2010</td>
<td>Mario Vargas Llosa</td>
<td>literature</td>
<td>Peru</td>
<td>Peru</td>
</tr>
</tbody>
</table>
Preview: Questions and answers

Which woman has won more than 1 Nobel prize?

(Marie Curie)

- **Question Answering (QA) and Information Retrieval (IR) involve many of the same principles.**
Aside – Question answering

How much potassium is in 450,000 cubic kilometers of bananas?

Input interpretation:
- banana
- amount
- 450 000 km$^3$ (cubic kilometers)
- potassium

Result:
$1.5 \times 10^{12}$ t (metric tons)
Answer questioning?

- Retrieving information can be a clever combination of many very simple concepts and algorithms.
Automatic summarization

Russia fights Napoleon and a Natalia likes Boris.

Don’t sit on a wall if you’re an egg.

Gregor turns into a bug.

Girl kills a woman, steals her shoes, then kills her sister.
Overview: NLC

• Is natural language computing (the discipline) hard?
  • Yes, because natural language
    • is highly ambiguous at all levels,
    • is complex and subtle,
    • is fuzzy and probabilistic,
    • involves real-world reasoning.
  • No, because computer science
    • gives us many powerful statistical techniques,
    • allows us to break the challenges down into more manageable features.
NLC in industry
My research
My research
My research

Please contribute your data here:

https://www.cs.toronto.edu/talk2me
Natural language computing

- **Instructor**: Frank Rudzicz (frank@cdf)
- **TAs**: Stefania R., Erin G., Louis T., Patricia T., Hamed H., Alex L., Arvie F., Hengwei G., Mengye R.
- **Meetings**: Mondays/Wednesdays 10h-11h in OI G162
- **Languages**: English, Python, Matlab.
- **Website**: [http://www.cs.toronto.edu/~frank/csc401/](http://www.cs.toronto.edu/~frank/csc401/)
- **You**: Understand basic **probability**, can **program**, or can pick these up as we go.
- **Syllabus**: Key **theory** and **methods** in statistical natural language computing. Focus will be on **Markov models**, **machine translation**, and **speech recognition**.
Office hours

• **Time:**
  • Immediately after lecture.

• **Location:**
  • BA7200; the streets

• **Ad hoc meetings:**
  • My main office is at 550 University.
Evaluation policies

- **General**: Three assignments: 20% each
  Final exam: 40%

- **Lateness**: 10% deduction applied to electronic submissions that are 1 minute late.
  Additional 10% applied every 24 hours up to 72 hours total, at which point grade is zero.

- **Final**: If you fail the final exam, then you fail the course.

- **Ethics**: Plagiarism and unauthorized collaboration can result in a grade of zero on the homework, failure of the course, or suspension from the University. See the course website.
Expansion

We need to pair you off

Enrollment
Grab your partner

1) Find a partner with whom you’d like to work.
   a. Feel free to use the CDF forum for this purpose.
2) Each of you create a file called ‘~/.csc401partner’.
3) Put the CDF user ID of your partner in that file.
4) If you really want to work alone, put your own CDF user ID in that file, and prepare for a semester of pain.
5) Do this by 18 January at 10h EST.

Teams will be assigned a random ‘Team Number’ that you will use to access specific data for the assignments, and...
IBM Watson

...access IBM Watson on BlueMix (https://console.ng.bluemix.net/)
Assignments

- Programming-intensive (and more work than previous years).
  - Languages: Python and MATLAB (opt. C/C++ modules).

- Subjects:
  - Assignment 1: Corpus statistics, sentiment analysis with Twitter Statistical techniques and classification.
  - Assignment 2: Statistical machine translation Statistical $n$-grams, smoothing, and multilingual word alignment.
  - Assignment 3: Automatic speech recognition Signal processing, phonetics, and hidden Markov models.
Assignment 1 – Sentiment analysis

• Involves:
  • Working with social media data (i.e., gathering statistics on some data from Twitter),
  • Part-of-speech tagging (more on this later),
  • Classification (e.g., SVM and decision trees in WEKA).
• Announcements: CDF forum, email, and @SPOClab.
• You should get an early start.
Projects – graduate students only

• Graduate students can **optionally** undertake a full-term **project** worth **60%** of their grade **instead** of the assignments.
  • Good for those, e.g., who prefer to work in teams. You might even get a **publication**!

• Teams must consist of **1 or 2 humans** *(no more, no fewer)*.
• Projects must contain a significant **programming** and **scientific** component.
• Projects must be **relevant** to the course.
Projects – graduate students only

• Some possible ideas for projects include:
  • A novel speech-to-speech machine translation.
  • A novel method of using data in language $A$ to train a classification system in language $B$ for $A \neq B$.

• If you decide to take this option, you have to notify me by email about your team by **18 January**!

• You will need to periodically submit **checkpoints** that build on their antecedents.
  • See course webpage for detailed requirements!
Reading

Mandatory (and FREE online!)

http://cognet.mit.edu/library/books/view?isbn=0262133601

Optional

SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

DANIEL JURAFSKY & JAMES H. MARTIN
Stats from last two years

Overall mark distribution

CSC401/2511 – Spring 2016
Stats from last year

The average overall grade among undergraduates was 66.2% \((\sigma=22.2)\). The average overall grade among graduates was 83.0% \((\sigma=7.5)\).

The grade range breakdown is:

<table>
<thead>
<tr>
<th>Grade</th>
<th>Undergrads</th>
<th>Grads</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

C+     | A-       |
Assignment 1 and reading

• **Assignment 1** available now (on course webpage)!
  • Due 12 February
  • **TAS:** Stefania Raimondo (**c3raimon@cdf**) ; Erin Grant (**t5grante@cdf**) .

• **Reading:**
  • Manning & Schütze:  Sections 1.3—1.4.2, Sections 6.0—6.2.1.

• **Wednesday 13 January:**
  • I’ll be in **Jamaica**. We’ll have a **tutorial** instead.