A New CAPTCHA Approach

**To complete your web registration, please prove that you’re human:**

When Littlefoot’s mother died in the original Land Before Time, did you feel sad?

- Yes
- No

(Bots: No living)

Administrivia

http://www.cs.pomona.edu/classes/cs159/
- Office hours, schedule, assigned readings, assignments
- Everything will be posted there

Read the “administrivia” handout!
- ~7 assignments (in a variety of languages)
- 4 quizzes
- Final project for the last 3-4 weeks
- Teams of 2-3 people
- Class participation
- Readings

Academic Honesty Collaboration

Administrivia

Assignment 0 posted already
- Shouldn’t take too long
- Due Thursday by 11:59pm (your time zone)

Assignment 1 posted
- Won’t cover all material until Thursday
- Due Wednesday 9/2 (11:59 – your time zone)
What to expect…

This course will be challenging for many of you
- assignments will be non-trivial
- content can be challenging

But it is a fun field!

We’ll cover
- basic linguistics
- probability
- the common problems
- many techniques and algorithms
- common machine learning techniques
- some recent advances in neural networks for language processing
- NLP applications

Requirements and goals

Requirements
- Competent programmer
- Some assignments in Java, but I will allow/encourage other languages after the first few assignments
- Comfortable with mathematical thinking
- We’ll use a fair amount of probability, which I will review
- Other basic concepts, like logs, summation, etc.
- Data structures
- trees, hashables, etc.

Goals
- Learn the problems and techniques of NLP
- Build real NLP tools
- Understand what the current research problems are in the field

Getting to know eachother

Videos before class
What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia

What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language…

- The book

Key: Natural text

“A growing number of businesses are making Facebook an indispensable part of hanging out their shingles. Small businesses are using …”

Natural text is written by people, generally for people

Why do we even care about natural text in computer science?

Why do we need computers for dealing with natural text?

Google’s search knows about over 130 trillion pages

In less than four years, Google’s search knowledge of pages has grown by more than 130 trillion new pages.

https://searchengineland.com/google-search-index-reached-130-trillion-pages-documents-563378
Web is just the start...

- e-mail
  - ~200-300 million e-mails a day
- corporate databases
- Twitter
  - ~500 million tweets a day
- Blogs
  - ~200 million different blogs
- Facebook

Why is NLP hard?

Iraqi Head Seeks Arms
Juvenile Court to Try Shooting Defendant
Stolen Painting Found by Tree
Kids Make Nutritious Snacks
Local HS Dropouts Cut in Half
Obesity Study Looks for Larger Test Group
British Left Waffles on Falkland Islands
Red Tape Holds Up New Bridges
Hospitals Are Sued by 7 Foot Doctors

Why is NLP hard?

User: Where is Escape Room playing in the Claremont Area?
System: Escape Room is playing at the Edwards in La Verne.
User: When is it playing there?
System: It's playing at 2pm, 5pm and 8pm
User: I'd like 1 adult and 2 children for the first show. How much would that cost?

Why is NLP hard?

Natural language:
- is highly ambiguous at many different levels
- is complex and contains subtle use of context to convey meaning
- is probabilistic?
- involves reasoning about the world
- is highly social
- is a key part in how people interact

However, some NLP problems can be surprisingly easy
**Different levels of NLP**

- pragmatics/discourse: how does the context affect the interpretation?
- semantics: what does it mean?
- syntax: phrases, how do words interact
- words: morphology, classes of words

**NLP problems and applications**

- What are some places where you have seen NLP used?
- What are NLP problems?

Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

I would’ve liked Dr. Dave to finish early. But he didn’t.
NLP problems and applications

Lots of problems of varying difficulty

Easier

- Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  
  I would’ve liked Dr. Dave to finish early. But he didn’t.
- Language identification
  
  Soy un maestro con queso.

Easier continued

- Truecasing
  
  i would’ve liked dr. dave to finish early. but he didn’t.
- Spell checking
  
  Identifying mispellings is challenging especially in the dessert.
- OCR

Moderately difficult

- Morphological analysis/stemming

  smarter
  smarter
  smart
  smartest

- Speech recognition

- Text classification

  SPAM

  Sentiment analysis

Moderately difficult continued

- Text segmentation: break up text by topics
- Part of speech tagging (and inducing word classes)
- Parsing

I eat sushi with tuna
NLP problems and applications

Moderately difficult continued

- word sense disambiguation
  
  As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

- grammar correction
  
  We are good at grammar.

- speech synthesis

NLP problems and applications

Hard (many of these contain many smaller problems)

Machine translation

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an email from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

NLP problems and applications

Information extraction

IBM hired Fred Smith as president.

<table>
<thead>
<tr>
<th>person</th>
<th>company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Smith</td>
<td>IBM</td>
<td>president</td>
</tr>
</tbody>
</table>
NLP problems and applications

Natural language understanding
- Text => semantic representation (e.g., logic, probabilistic relationships)

Information retrieval and question answering
- "How many programmers in the child care department make over $50,000?"
- "Who was the fourteenth president?"
- "How did he die?"

Text simplification
Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Where are we now?
Many of the "easy" and "medium" problems have reasonable solutions
- spell checkers
- sentence splitters
- word segmenters/tokenizers

Where are we now?
Parsing
- Stanford Parser (http://nlp.stanford.edu:8080/parser/)
Where are we now?

Machine translation

How is it?

- Getting better every year
- Enough to get the jist of most content, but still no where near a human translation
- Better for some types of text

*http://translate.google.com*

Many commercial versions...
- Systran
- Language Weaver

Where are we now?

Information extraction

- Structured documents (very good!)
  - www.dealtime.com
  - www.google.com/shopping
- ART technologies
  - Lots of these
  - FlipDog
  - WhizBang! Labs
  - ... (work fairly well)

Where are we now?

CMU’s NELL (Never Ending Language Learner)

*http://rtw.ml.cmu.edu/rtw/*
Where are we now?

Information retrieval/query answering

- search engines:
  - pretty good for some things
  - does mostly pattern matching and ranking
    - no deep understanding
    - still requires user to “find” the answer

Wolfram Alpha
Where are we now?

Question answering:

- Many others systems
- TREC question answering competition
- language computer corp
- answerbus
- ...

Summary:

NewsBlaster
(Columbia) http://newsblaster.cs.columbia.edu/

-Ukraine crisis: Russian forces reportedly seen in rebel cities Donetsk and Luhansk

Summary from multiple countries, from articles in English

- Pro-Russian rebels fulfill their demand for full independence Monday, saying they would not be able to keep their autonomy in the eastern part of Russia. Article 15:
- The Ukrainian government, which has been struggling to maintain its hold on the region, has refused to negotiate with the rebels. Article 16:
- A new round of talks is due to begin Tuesday in Minsk, the capital of Belarus, to discuss the crisis in the eastern part of the country. Article 17:

-Pro-Russian rebels have occupied the town of Slavyansk, Ukraine’s second-largest city, in recent days. Article 18:
Where are we now?

Voice recognition
- pretty good, particularly with speaker training
- Apple OS/Siri
- Android/Google
- Alexa, Google Assistant, etc...
- IBM ViaVoice
- Dragon Naturally Speaking

Speech generation
- The systems can generate the words, but getting the subtle nuances right is still tricky
- Apple OS
- http://translate.google.com

Other problems

Many problems untackled/undiscovered!