15383: Intro to Text Processing

Behrang Mohit

Welcome!

• Welcome To 15383!

• Instructor: Behrang Mohit
  – Office: 1004
  – first-name@cmu.edu
The AI Dream

• Creating intelligent systems capable of simulating humans

Language and Text

• Has been present since early days of human civilization.
21st Century: So Much Text!

• Problem: Information overload!

21st Century: So Much Text!

• Exponential growth of text in the *surface* web and also the *deep* web.
  – 400m tweets/day
Generate, Organize and Process

- Need to generate, organize and process text:
  - Different topics and genres
    - News, science, sport, film subtitles, children stories, jokes,...
  - Different languages
  - Different platforms and mediums
    - prints, desktop, mobile device, TV, ...
    - Internet
      - Official channels (government and corporate webpages)
      - Personal pages, social media

Examples of Text Processing Tasks

- Searching and categorizing
- Extracting information from text
  - Who is doing what to whom when
- Summarize text and answer questions
- Translate
- Understand text
- Chat and counsel humans (psychotherapy)
15383: What Will We Learn Here?

- Features of text: Corpus, Encoding
- Basics of statistics
- Text Organization:
  - Document Classification
  - Search: Information Retrieval
- Linguistics in text processing
  - a.k.a. Natural Language Processing
- Python programming (NLTK)
- Course project: Sentiment Analysis on Social Media
Text As A Medium

- Natural Language Processing
  - Speech
  - Text
    - Focus on natural language text

Today

- An overview of topics
Text And Its Encodings

• Text: Words, sentences, documents, ...
• Processing and organizing large volumes of text
  – Corpus (corpora)
    • For building and evaluating text processing systems
    • Might include extra linguistics information
• Encoding the text

Statistics In Text Processing

• Rule-based systems vs. statistical systems
• Probabilities
• Statistical learning
  – Supervised learning
Text Organization

- Large volumes of text → organized text
- Document classification
  - Sport, politics, science, ...
  - Email classification
    • Work, Fun, Spam, ...

- Searching documents
  - Ask, Google, Bing, etc.

Natural Language Processing is ...

- NLP or
  - Computational Linguistics
  - Human Language Technologies

- Goal: Making computers capable of using human language as their input or output, performing intelligent tasks.
NLP and Artificial Intelligence

• NLP is the fundamental problem of Artificial Intelligence (AI).

• Turing test for the intelligence of a machine
  – If a human judge can not distinguish between a machine and human in a conversation framework, the machine passes the Turing test.

The Dream of Talking Machines!

• 2001 Space Odyssey
  – Dave (human): Open the door Hal
  – HAL (machine): I’m sorry Dave, I can’t do that.

• HAL: An intelligent system capable of:
  – Understanding and generating human language
Linguistic Layers

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

Linguistics Layers: Morphology

- What are building blocks of words?
  - goes ➔ go + es
  - prettiest ➔ Pretty + est
- Different levels of complication in morphology
  - English
  - Arabic, Finnish, Turkish
    - wysyaktobun ➔ w + s + yaktob + un
    - And will write they ➔ and they will write
Linguistic Layers: Syntax

- How do words come together to form more complex units?
  - Phrases, sentences, relationship between phrases
  - Mostly at the sentence level
  - Zeinab bought a book.
    - Noun Verb Det Noun Punctuation
    - Subject Verb Object

Linguistic Layers: Semantics

- What is the meaning of terms in a sentence
  - Suhail bought a book.
    - Commercial transaction:
      - Buyer: Suhail
      - Action: buying
      - Commodity: book
Linguistic Layers: Pragmatics and Discourse

• Going beyond a sentence-level analysis
  – Ahmad arrived in Doha. He was accompanied by his family. They went directly to a wedding from the airport.

Linguistics Layers

• Phonetics
• Phonology
• Morphology
• Syntax
• Semantics
• Pragmatics
• Discourse
Two Major NLP Challenges

• Challenge 1: Getting the proper linguistic representation of the input
  – From sound waves to text
  – From the sentence to syntactic tree
    • Mariem bought a book ➔ Mariem: sub, book: Obj
  – ....

Two Major NLP Challenges

• Challenge 2: Ambiguity in language

• A language understanding example
  – “At last, a computer that understands you like your mother!”
    • Ad from Microsoft (in early 1980s)
    • Example by Stuart Shiebert
A Computer That Understands...

• At last, a computer that understands you like your mother!

• Computer understands you as well as your mother understands you
  • Computer understands that you like your mother
  • Computer understands you as well as it understands your mother

• Problem: Ambiguity in human expressions
A Computer That Understands...

• At last, a computer that understands you like your mother!

• Computer understands you as well as your mother understands you

• Computer understands that you like your mother

• Computer understands you as well as it understands your mother

• Problem: Ambiguity in human expressions

The Ambiguity Problem

• Humans use common-sense, bits of culture, world knowledge in their expressions!
  – Do computers understand all of those?
Levels of Ambiguity: Acoustics

- Understands you like your mother
- Understands you **lie cured** mother

- It is hard to *recognize speech*.
- It is hard to *wreck a nice beach*.

Levels of Ambiguity: Syntax

- Different sentence structure (syntax):
  
  – Computer that understands you *(like your mother [does])*  
  – Computer that understand *(that] you like your mother)*
Levels of Ambiguity: Semantics

• ... knows you like your **mother**
  • The female parent
    – Most probably
  • A vat (dish) for making vinegar

• We put our money in the **bank**
  – Money bury under the mud (river bank)!
  – Financial institution
    • Most probably

Levels of Ambiguity: Discourse

• Leila says they are selling a computer that knows you like your mother. But **she** doesn’t seem to be happy about it.
  
  – Who does **she** refer to?
    • mother, computer, Leila?
Let’s Disambiguate

• I saw her duck with a telescope

• I used a telescope to see her duck
  • I saw her duck that was carrying a telescope.
  • I used a telescope to see her ducking
  • I saw her ducking using a telescope
  • I cut her duck with a telescope
  • ....
Let’s Disambiguate

• I saw her duck with a telescope

• I used a telescope to see her duck
• I saw her duck that was carrying a telescope.
  • I used a telescope to see her ducking
  • I saw her ducking using a telescope
  • I cut her duck with a telescope
  • ....
Let’s Disambiguate

• I saw her duck with a telescope

• I used a telescope to see her duck
• I saw her duck that was carrying a telescope.
• I used a telescope to see her ducking
• I saw her ducking using a telescope
• I cut her duck with a telescope
• ....

Linguistics Layers

• Phonetics
• Phonology
• **Morphology**
• Syntax
• **Semantics**
• Pragmatics
• Discourse
Course Plan

1. Text
2. Text organization
3. Processing with three linguistic layers
   – Words (morphology), Syntax and Semantics
4. NLP applications

Application: Spell Checking

• Ali arrived at scool
  – scull
  – school
  – cool
  – spool

• Idea: Look at the previous words to decide between the given correct options.
  – Use statistics
    • $Pr(\text{arrived at } \Rightarrow \text{ school})$
    • $Pr(\text{arrived at cool})$
    • $Pr(...)$
Application: Named Entity Recognition

- Names of Persons, Locations, Organization, ...

- George Washington ruled America for two terms.
- George Washington University announced ...
- As George was walking in Washington, he ...

Application: Text summarization

- Summarizing large volumes of text
  - Locate the important parts of the text and form sentences with them.
    - Natural language generation
  - Useful for governments, companies, etc.

- Word Processing and browser offer the service
Application: Machine Translation

- Text translation from one language to another
  - Dealing with differences in two languages
    - English: Subject-verb-object
    - Arabic: Verb Subject Object
  - Ambiguities in two languages
  - Semantic differences:
    - Concept of cousin

Application: Sentiment Analysis

- Imagine
  - Your company (e.g. Apple) has released a new product (e.g. iphone) and wants estimate the initial reaction of customers
  - You’re campaigning for a politician and you want to estimate people’s reaction to his last night speech.
Application: Sentiment Analysis

• Distinguish between objective and subjective statements.
  – News vs. Opinion

• Find polarity of statements
  – Product reviews:
    • The new laptop is hot!
    • The new laptop gets very hot!

• Example: Organizing hundreds of film reviews
  – “This is a feel-good blockbuster production with an excellent technical setup.”
  – Bottom-line: Does this author likes the movie?

Positive, Negative or ...?

• You’ll see a tweet and will say if it’s
  – Subjective or Objective
    • Does it carry any opinion/sentiment?
  – If subjective: Positive, Negative or Neutral?
Positive, Negative or ...?

• “Authorities are only too aware that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but only a tenth of the distance from the Pakistani border, and are desperate to ensure instability or militancy does not leak over the frontiers.”

Positive, Negative or ...?

• Friday evening plans were great, but Saturday's plans didn't go as expected -- I went dancing & it was an ok club, but terribly crowded :-(

Positive, Negative or ...?

• Iran and 5+1 walk in to new rounds of negotiations on the nuke program. #iran #us

• “obama should be impeached on TREASON charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. #Coward #Traitor.”
Positive, Negative or ...?

- “My graduation speech: "I'd like to thanks Google, Wikipedia and my computer! :D #iThingteens”

- WHY DO YOU GUYS ALL HAVE MR. KENNEDY! HE’S A F...G DOUCHE!
Required knowledge

• Language:
  – Words: desperate, leak, impeach, TREASON
  – Sarcasm
    • My graduation speech: "I'd like to thanks Google, Wikipedia and my computer!
  – Social media’s non-formal and erroneous language
    • didn’t, :-( omg thanx gr8 b4

• Medium:
  – Hash tags in tweets

Problem: Analyzing emotions (Contd.)

• Mass analysis of linguistic emotions
  – On Social Networks
Communication

• Behrang Mohit
  – CMUQ 1004
  – *(behrang@cmu.edu)*

• Course webpage:
  – http://www.qatar.cmu.edu/~behrang/15383/

Administrative

• This is mostly a project-based course
  – Most of your grade is decided by the programming assignments and the **final project**.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance and Participation</td>
<td>5%</td>
</tr>
<tr>
<td>Python Labs</td>
<td>10%</td>
</tr>
<tr>
<td>Quizzes</td>
<td>10%</td>
</tr>
<tr>
<td>Programming assignments</td>
<td>30%</td>
</tr>
<tr>
<td>Final Project</td>
<td>45%</td>
</tr>
</tbody>
</table>
Resources

• Lecture notes
• Natural Language Processing with Python

• Further Reading: Speech and Language Processing
  – Three copies on reserve at the library.

NLTK

• Natural Language Toolkit
  – Python library for natural language processing

  – @library
  – Online: http://www.nltk.org/book