Contents

• Syllabus
• Homeworks 1 and 2
  – (do them before next week)
• Some intro
Details:

• 538: introductory level, no formal pre-requisites
• 438: LING 388 or familiarity with one or more of the following: formal languages, syntax, data structures, or compilers
• Instructor: Sandiway Fong, Depts. of Linguistics Office: Douglass 311
• Hours:
  – by appt. or walk-in
  – after class (best if you have quick Qs)
• Email: sandiway@email.arizona.edu
• Meet: Tues/Thurs in McClelland Park 102, 3:30-4:45pm
• No class on
  – Week of 18th-22nd Sept (out of town)
  – Nov 23rd (Thanksgiving)
Syllabus

• Course objectives:
  – introduction to computational linguistics
  – survey a range of topics
  – introduction to programming

• Expected learning outcomes:
  – acquire ability to write (at least, short) programs
  – familiarity with basic concepts, techniques and applications
  – be equipped to take more advanced classes in computational linguistics, e.g. 581 (Spring)
Syllabus

• Grading
  – 438
    • homeworks 100%
    • **note**: *all homeworks are required*
  – 538
    • homeworks 75%
    • *(homeworks may be a superset of the exercises for 438)*
    • chapter presentation 25%

• Homework submissions
  – email only
  – [sandiway@email.arizona.edu](mailto:sandiway@email.arizona.edu)
  – by midnight of due date
  – typically: one week
  – *(homeworks will be introduced and reviewed in class)*
Syllabus

• Homeworks
  – you may discuss questions with other students
  – however, you must write it up yourself (*in your own words, your own code etc.*)
  – cite (web) references and your classmates (*in the case of discussion*)
  – Student Code of Academic Integrity: *plagiarism* etc.
    • [http://deanofstudents.arizona.edu/codeofacademicintegrity](http://deanofstudents.arizona.edu/codeofacademicintegrity)

• Revisions to the syllabus
  – “the information contained in the course syllabus, other than the grade and absence policies, may be subject to change with reasonable advance notice, as deemed appropriate by the instructor.”
Syllabus

• Absences
  – *tell me ahead of time so we can make special arrangements*
  – *I expect you to attend lectures (though attendance will not be taken)*

• Required text
  – *Speech and Language Processing*, Jurafsky & Martin, 2\textsuperscript{nd} edition, Prentice Hall 2008

• Special equipment
  – *technically none, but a laptop is recommended*
  – all software required for the course is freely available off the net

• Classroom etiquette
  – ask questions
  – use your own laptop or lab computer

• Topics (16 weeks)
  – Programming Language: Perl (Python)
  – Regular Expressions
  – Automata (Finite State)
  – Transducers (Finite State)
  – Programming Language: Prolog (definite clause grammars)
  – Part of Speech Tagging
  – Stemming (Morphology)
  – Edit Distance (Spelling)
  – Grammars (Regular, Context-free)
  – Parsing (Syntax trees, algorithms)
  – N-grams (Probability, Smoothing)
  – *and more ...*
Course website

• Download lecture slides from my homepage
  – [http://elmo.sbs.arizona.edu/~sandiway/#courses](http://elmo.sbs.arizona.edu/~sandiway/#courses)
  – available from just before class time
    • (afterwards, look again for corrections/updates)
  – in .pptx (good for animations) and .pdf formats
Course website

About 18,400 results (0.81 seconds)

1st hit

Sandiway Fong
elmo.sbs.arizona.edu/sandiway/
Phone: 520 626 5657. Fax: 520 626 9014. Office: 311 Douglass. Email: sandiway at email dot arizona dot edu. About me: (Updated 5/2004) here. Resume ...

Sandiway Fong | The Department of Linguistics
https://linguistics.arizona.edu/user/sandiway-fong
About Sandiway Fong. Human Language Technology MA Program Coordinator: I am a computational linguist. I am interested in all aspects of language and ...

Sandiway Fong - u.arizona.edu
www.u.arizona.edu/~sandiway/
Miss a lecture?

- Lectures will be recorded using the panopto system
  - Accessible via the course webpage
  - Just use your browser
  - (video, laptop screen, synchronized slides, keyword search)
Textbook (J&M)

2008 (2nd edition)

Nearly 1000 pages

*(maybe more than a full year’s worth...)*

25 chapters

Divided into 5 parts

I. Words

II. Speech – *not this course*

III. Syntax

IV. Semantics and Pragmatics

V. Applications
1. Introduction
1.1. Knowledge in Speech and Language Processing
1.2. Ambiguity
1.3. Models and Algorithms
1.4. Language, Thought, and Understanding
1.5. The State of the Art
1.6. Some Brief History
1.6.1. Foundational Insights: 1940s and 1950s
1.6.2. The Two Camps: 1957–1970
1.6.5. The Field Comes Together: 1994–1999
1.6.7. On Multiple Discoveries
1.6.8. A Final Brief Note on Psychology
1.7. Summary
Bibliographical and Historical Notes
Syllabus

• Coverage
  – There are no formal prerequisites for 538
  – I’m gonna assume you don’t know how to program at all (yet)
    • we’re going to use Perl and sometimes Python
    • advice: learn both ...
  – Topics: selected chapters from J&M
    • Chapters 1–6, skip Speech part (7–11), 12–25
Chapter 1
Introduction

David Boorman: Open the pod bay doors, HAL.
HAL: I’m sorry Dave, I’m afraid I can’t do that.
Stanley Kubrick and Arthur C. Clarke, screenplay of 2001: A Space Odyssey

The idea of giving computers the ability to process human language is as old as the idea of computers themselves. This book is about the implementation and implications of that exciting idea. We introduce a vibrant interdisciplinary field with many names corresponding to its many facets, names like speech and language processing, human language technology, natural language processing, computational linguistics, and speech recognition and synthesis. The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.

One example of a useful such task is a conversational agent. The HAL 9000 computer in Stanley Kubrick’s film 2001: A Space Odyssey is one of the most recognizable characters in 20th century cinema. HAL is an artificial agent capable of such advanced language behavior as speaking and understanding English, and at a crucial moment in the plot, even reading lips. It is now clear that HAL’s creator, Arthur C. Clarke, was a little optimistic in predicting when an artificial agent such as HAL would be available. But just how far off was he? What would it take to create at least the language-related parts of HAL? We call programs like HAL that converse with humans in natural language conversational agents or dialogue systems. In this text we study the various components that make up modern conversational agents, including language input (automatic speech recognition and natural language understanding) and language output (dialogue and response planning and speech synthesis).

Let’s turn to another useful language-related task, that of making available to non-English-speaking readers the vast amount of scientific information on the Web in English. Or translating for English speakers the hundreds of millions of Web pages written in other languages like Chinese. The goal of machine translation is to automatically translate a document from one language to another. We introduce the algorithms and mathematical tools needed to understand how modern machine translation works. Machine translation is far from a solved problem; we cover the algorithms currently used in the field, as well as important component tasks.

Many other language processing tasks are also related to the Web. Another such task is Web-based question answering. This is a generalization of simple Web search, where instead of just typing keywords, a user might ask complete questions, ranging from easy to hard, like the following:

- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
Homework 2

• Install Perl (and Python)
  – could be the null homework ...
Homework: Install Perl

• Install Perl on your laptop
  – should be pre-installed on macs and Linux (Ubuntu), check your machine
  – on Windows PCs, if you don’t already have it, it’s freely available here
  – [www.perl.org](http://www.perl.org)
Homework: Install Perl

- Ubuntu (Terminal):
  ```
  perl -v
  perl
  ```

- Mac (Terminal): *(complete path specified) /usr/bin/perl*
  ```
  Perl may be copied only under the terms of either the Artistic License or the GNU General Public License, which may be found in the Perl 5 source kit.
  Complete documentation for Perl, including FAQ lists, should be found on this system using "man perl" or "perldoc perl". If you have access to the Internet, point your browser at http://www.perl.org/, the Perl Home Page.
Homework: Install Perl

• Actually, on my Macbook ...

```
[dhcp-10-142-182-240:~ sandiway$ perl -v

This is perl 5, version 24, subversion 1 (v5.24.1) built for darwin-thread-multi-2level

Copyright 1987-2017, Larry Wall

Perl may be copied only under the terms of either the Artistic License or the GNU General Public License, which may be found in the Perl 5 source kit.

Complete documentation for Perl, including FAQ lists, should be found on this system using "man perl" or "perldoc perl". If you have access to the Internet, point your browser at http://www.perl.org/, the Perl Home Page.

[dhcp-10-142-182-240:~ sandiway$ which perl
/opt/local/bin/perl
dhcp-10-142-182-240:~ sandiway$ ]

commands:
perl -v
perl -V
```
Homework: Install Perl

https://www.perl.org/learn.html

Get Started
- learn.perl.org
- A brief introduction
- Free online Perl books
- Join your local community
- Books and More

Modules
If you are looking for a list of recommended modules for many day-to-day tasks, look at Task::Kensho.
Learning Perl

- Learn Perl
  - Books...
  - Online resources
Homework: Install Python

```python
# Python 3: Simple arithmetic
>>> 1 / 2
0.5
>>> 2 ** 3
8
>>> 17 / 3 # classic division returns a float
5.666666666666667
>>> 17 // 3 # floor division
5
```

Intuitive Interpretation

Calculations are simple with Python, and expression syntax is straightforward: the operators +, - , * , and / work as expected; parentheses () can be used for grouping. More about simple math functions in Python 3.
Homework: Install Python

- Not quite backwards compatible!
- See [https://docs.python.org/3/whatsnew/3.0.html](https://docs.python.org/3/whatsnew/3.0.html)
Homework: Install Python

- On my Macbook, I have both:

```
[dhcp-10-142-182-240:~] sandiway$ which python
/usr/bin/python

[sandiway$] python
Python 2.7.10 (default, Feb 7 2017, 00:08:15)
[GCC 4.2.1 Compatible Apple LLVM 8.0.0 (clang-800.0.34)] on darwin
Type "help", "copyright", "credits" or "license" for more information.

>>> 
```

```
[dhcp-10-142-182-240:~] sandiway$ which python3
/Library/Frameworks/Python.framework/Versions/3.5/bin/python3

[sandiway$] python3
Python 3.5.2 (v3.5.2:4def2a2901a5, Jun 26 2016, 10:47:25)
[GCC 4.2.1 (Apple Inc. build 5666) (dot 3)] on darwin
Type "help", "copyright", "credits" or "license" for more information.

>>> 
```
Which one is easier?

• A subjective question ...
• All good programmers know more than one programming language
• In NLP, Python is more popular (you will acquire this one anyway), so let’s do Perl plus some Python ...
• https://www.tecmint.com/python-vs-perl-debate-what-should-i-learn-python-or-perl/
Which one is friendlier?

• let’s see...
Language and Computers

• Enormous amounts of data stored
  – world-wide web (WWW)
  – corporate databases
  – Dark Web
  – your own SSD or hard drive

• Major categories of data
  – numeric
  – **Language**: words, text, sound
  – pictures, video
Language and Computers

• We know what we want from computer software
• “killer applications”
  – those that can make sense of language data
    • retrieve language data: (IR)
    • summarize knowledge contained in language data
    • sentiment analysis from online product reviews
    • answer questions (QA), make logical inferences
    • translate from one language into another
    • recognize speech: transcribe
    • etc...
Language and Computers

• In other words, we’d like computers to be smart about language
  • possess “intelligence”
  • pass the Turing Test ...
  • but not be too smart?

From 2001...

HAL


We Should Be as Scared of Artificial Intelligence as Elon Musk Is

Steven Finlay
Aug 18, 2017

Elon Musk recently commented on Twitter (TWTR,+0.75%) that artificial intelligence (AI) is more dangerous than North Korea. It’s not the first time that the entrepreneur has warned about the dangers of AI. Should we all be afraid as he is? Will AI lead to a huge disaster or robot takeover that destroys humanity?

Language and Computers

• (Un)fortunately, we’re not quite there yet...
  – still gap between what computers can do and what we want them to be able to do

Often quoted (but never verified):

"The spirit is strong, but the flesh is weak" was translated into Russian and then back to English, the result was "The vodka is good, but the meat is rotten."

but with Google translate or babelfish, it’s not difficult to find (funny) examples...
Language and Computers

• and how can we tell if the translation is right anyway?

• http://fun.drno.de/pics/english/only-in-china/TranslateServerError.jpg
Applications

- *technology is still in development*

• even if we are willing to pay...
  - machine translation has been worked on since after World War II (1950s)
  - still not perfected today
  - *why?*
  - what are the properties of human languages that make it hard?
Language and Computers

Recursive nature of language ...

Biden apologizes to Obama for marriage controversy
From Jessica Yellin, CNN Chief White House Correspondent
updated 10:20 PM EDT. Thu May 10, 2012

President Obama explains his position on same-sex marriage
Language and Computers

• Obama: "At a certain point, I've just concluded that for me personally it is important for me to go ahead and affirm that I think same-sex couples should be able to get married."

Is this sentence complicated? Why?
It's cool that computer programs can diagram this sentence!
Obama: ‘I Think Same-Sex Couples Should Be Able to Get Married’
Language and Computers

• Obama: "At a certain point, I've just concluded that for me personally it is important for me to go ahead and affirm that I think same-sex couples should be able to get married."

Most summarizer programs can't do this ...
Language and Computers

http://tomato.banatao.berkeley.edu:8080/parser/parser.html
Natural Language Properties

• *which properties are going to be difficult for computers to deal with?*

• **grammar** (Rules for putting words together into sentences)
  - *How many rules are there?*
    • 100, 1000, 10000, more …
  - Portions learnt or innate
  - *Do we have all the rules written down somewhere?*

• **lexicon** (Dictionary)
  - How many words do we need to know?
    • 1000, 10000, 100000 …

• **meaning and inference** (semantic interpretation, commonsense world knowledge)
Computers vs. Humans

• Knowledge of language
  – Computers are way faster than humans
    • *They kill us at arithmetic*
    • *and chess*
  
  *and Jeopardy as well ...*

  **AI Apocalypse**

  – But human beings are so good at language, we often take our ability for granted
    • *Processed without conscious thought*
    • *Do pretty complex things*
Examples

• Ungrammaticality
  – *Which book did you file the report without reading?
  – vs. which book did you file without reading?
  – * = ungrammatical

  Colorless green ideas sleep furiously.
  Furiously sleep ideas green colorless.

  – ungrammatical vs. incomprehensible
Examples

• interpretation
  • John is too stubborn to talk to ___
  • John is too stubborn to talk to Bill

• Subject of "to talk to"

  John is too stubborn [ ] to talk to [ ]
  John is too stubborn [ ] to talk to Bill
Examples

• Ambiguity
  – where can I see the bus stop?
  
  – *stop*: (1) verb, or (2) part of the noun-noun compound *bus stop*
  
  – Context (Discourse or situation)