LING/C SC/PSYC 438/538

Lecture 1
Sandiway Fong
Syllabus

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 and Computer Science

• Office: Douglass 311

• Hours:
  – by appt. or walk-in
  – after class (best if you have quick Qs)

• Email: sandiway@email.arizona.edu

• Meet: Mondays/Wednesdays in Shantz 338, 4:45-6pm

• No class on
  – Monday September 7th (Labor Day)
  – Wednesday November 11th (Veterans Day)
  – Week after September 11th (out of town), plus Monday 21st
  – Monday October 12th
Syllabus

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

• Expected learning outcomes:
  – acquire ability to write 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, will be a superset of the exercises for 438)*
    • chapter presentation 25%

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

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

• Special equipment
  – none
  – 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
  – 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://dingo.sbs.arizona.edu/~sandiway/#courses](http://dingo.sbs.arizona.edu/~sandiway/#courses)
  – available from just before class time
    • (afterwards, look for corrections/updates)
  – in .pptx (animations) and .pdf formats
Course website

About 6,120 results (0.41 seconds)

Sandiway Fong
dingo.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
www.u.arizona.edu/~sandiway/^
Sandiway Fong ... http://dingo.sbs.arizona.edu/~sandiway. [Note: this has been ...
Software releases, e.g. wnconnect, sandiway sudoku can be found there as well.

Sandiway Fong | The Department of Linguistics
linguistics.arizona.edu/user/sandiway-fong
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
Book chapters

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 don't assume you know how to program (yet)
    • we’re going to use Perl
    • *Python is another (perhaps more) popular language*

– Topics: selected chapters from J&M
  • Chapters 1–6, skip Speech part (7–11), 12–25
Homework: Reading

• Chapter 1 from JM
  – introduction and history
  – available online
  – [Link to online resource]

  - Chapter 1 from J.M.

• Whole book is available as an e-book
  – [Link to e-book]

Chapter 1
Introduction

Dave Bowman: 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, sourceplay 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-machine 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: 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
  - [http://www.activestate.com/](http://www.activestate.com/) *(don’t pay, get the free version)*
Homework: Install Perl

• Ubuntu (Terminal):

```
sandiway@sandiway-VirtualBox:~
sandiway@sandiway-VirtualBox:~$ perl -v
This is perl 5, version 18, subversion 2 (v5.18.2) built for x86_64-linux-gnu-thread-multi
(with 41 registered patches, see perl -V for more detail)
Copyright 1987-2013, Larry Wall
```

• Mac (Terminal):

```
Last login: Sun Aug 23 02:05:35 on console
dhcp-10-142-149-113:$ sandiway$ perl -v
This is perl 5, version 16, subversion 3 (v5.16.3) built for darwin-thread-multi-2level
Copyright 1987-2012, Larry Wall
```
Homework: Install Perl

Other methods
See http://learn.perl.org/installing/
Learning Perl

• Learn Perl
  – Books...
  – Online resources
    • [http://learn.perl.org/](http://learn.perl.org/)
    • Next time, we begin with ...
    • [http://perldoc.perl.org/perlintro.html](http://perldoc.perl.org/perlintro.html)
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 ...
Language and Computers

• In other words, we’d like computers to be smart about language
  — possess intelligence
  — well, perhaps not too smart…

5 Very Smart People Who Think Artificial Intelligence Could Bring the Apocalypse

Professor Stephen Hawking is another worrier, saying in December that “the primitive forms of artificial intelligence we already have, have proved very useful. But I think the development of full artificial intelligence could spell the end of the human race.”

"AI Apocalypse"

From 2001... (HAL)
Language and Computers

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

Often quoted (but not 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
Language and Computers

Do you trust Google Translate?

- a real case: 4,000,000 yen or 40,000 yen?

![Image of a violin with a Google Translate tool showing the conversion from 50,404.496 US dollars to 4000000 Japanese yen.]

<table>
<thead>
<tr>
<th>Product Specifications Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Applicable Models</strong></td>
</tr>
<tr>
<td>Maikurofosumaunto</td>
</tr>
</tbody>
</table>

For electrically connect with the lens, AF from photographing the diorama-like shift in the direction of tilt up and down vertically with auto exposure, depending on how it can be taken as the expensive Chirutoshifutorenzu All Four Thirds lenses can be used as a Chirutoshifutorenzu (500g recommends the following lenses. by the lens may have a material unsuitable for practical use)
Language and Computers

- Puzzle: translation of 4万円以下

Original Japanese text: 4万円以下

Determined (less than 4 million yen)

4万円以下

10,000 yen

以下

less than/below/not exceeding

4万円以下

40,000 yen less

with auto-detect on
Language and Computers

• Non-compositionality Puzzle
What happened? 4万円以下 – can be segmented as follows:

- 4 万円 10,000 yen
- 以下 less than/below/not exceeding
- 万円以下 Million yen
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 ...
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?
Language and Computers

It's cool that computer programs can diagram this sentence!
A Sports Shooter Shoots Shooters Shooting Sports
A sports shooter shoots shooters shooting sports
Language and Computers

May 9, 2012 2:57pm

Obama: ‘I Think Same-Sex Couples Should Be Able to Get Married’

 Executive Summarization
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 ...
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

• Knowledge
  Which report did you file ___ without ___ reading ___?
  – (Parasitic gap sentence)

We take for granted this process of “filling in” or recovering the missing information
Examples

• Ungrammaticality
  – *Which book did you file the report without reading?

  – * = ungrammatical

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

  – ungrammatical vs. incomprehensible
Examples

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

• Subject of "to talk to"

<table>
<thead>
<tr>
<th>John is too stubborn</th>
<th>to talk to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(fill in)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>John is too stubborn</th>
<th>to talk to Bill</th>
</tr>
</thead>
</table>
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)
Examples

• The human parser has quirks
  • Ian told the man that he hired a story
  • Ian told the man that he hired a secretary

  • Garden-pathing: a temporary ambiguity
  • *tell*: someone something vs. ...

• More subtle differences
  • The reporter who the senator attacked admitted the error
  • The reporter who attacked the senator admitted the error