Computational Linguistics 1

CMSC/LING 723, LBSC 744

Kristy Hollingshead Seitz
Institute for Advanced Computer Studies
University of Maryland

1 September 2011

Agenda

- Administrivia
- Introduction to Computational Linguistics & applications
- Rule-based & statistical NLP

Administrivia

- Course webpage:
  `www.umiacs.umd.edu/~hollingk/classes/CompLing1-f11.html`
- Course mailing list:
  `umd-cmsc723-fall-2011@googlegroups.com`
- Textbook
  - Speech and Language Processing,
    Daniel Jurafsky and James H. Martin
- Teaching Assistant: Alex Ecins
  - Office hours

Course Policies

- Policies
  - Attendance
  - Homework
    - Submitted by e-mail to: `compling723.fall2011@gmail.com`
    - Computer access?
  - Late/incomplete work
  - Exams
- Grading
  - Exams
  - Homeworks
  - In-class participation
    - Readings

Pre-requisites

- Must have strong computational background
- Be a competent programmer
  - Depth-first search
  - Programming language: recommend Python/NLTK
- Be interested in linguistics
  - "The aged bottle flies fast"

- Enrollment/waitlist
- Machine Learning students?

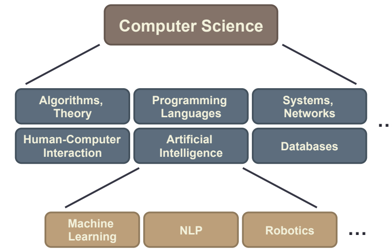
Agenda

- Administrivia
- Introduction to Computational Linguistics & applications
- Rule-based & statistical NLP

What is Computational Linguistics?

- Computer processing of naturally-occurring language
  - What humans do when processing language
  - (vs) What linguists do when processing language
- Various names
  - Computational linguistics
  - Natural language processing (NLP)
  - Speech/language/text processing
  - Human language technology
- Interdisciplinary field
  - Roots in linguistics and computer science (specifically, AI)
  - Influenced by electrical engineering, cognitive science, psychology, and other fields
  - Dominated today by machine learning and statistics

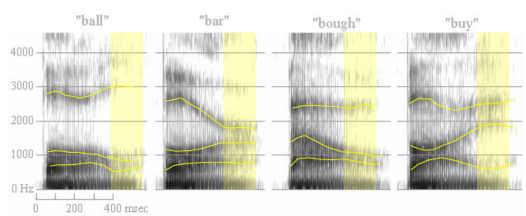
Where does NLP fit in CS?



from Jimmy Lin

Applications

- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)



Applications

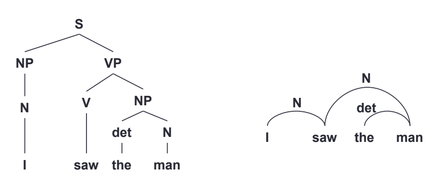
- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)
- Optical Character Recognition (OCR)
  - Image processing. e.g., captchas

Applications

- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)
- Optical Character Recognition (OCR)
  - Image processing. e.g., captchas
- Parsing: syntax & semantics
  - "The aged bottle flies fast"

Syntactic Analysis

- Parsing: the process of assigning syntactic structure



$[_S\ [_{NP}\ I\ ]\ [_{VP}\ saw\ [_{NP}\ the\ man]\ ]\ ]$

Semantics

- Different structures, same* meaning:
  - I saw the man.
  - The man was seen by me.
  - The man was who I saw.
  - …
- Semantic representations attempt to abstract “meaning”
  - First-order predicate logic:
    ∃ x, MAN(x) ∧ SEE(x, I) ∧ TENSE(past)
  - Semantic frames and roles:
    (PREDICATE = see, EXPERIENCER = I, PATIENT = man)

Lexical Semantics

- Any verb can add “able” to form an adjective.
  - I taught the class. The class is teachable.
  - I loved that bear. The bear is loveable.
  - I rejected the idea. The idea is rejectable.
- Association of words with specific semantic forms
  - John: noun, masculine, proper
  - the boys: noun, masculine, plural, human
  - load/smear verbs: specific restrictions on subjects and objects

Applications

- Speech recognition and synthesis
  - Lots of signal processing to go from raw waveforms into text (and vice versa)
- Optical Character Recognition (OCR)
  - Image processing. e.g., captchas
- Parsing: syntax & semantics
  - "The aged bottle flies fast"
- Machine translation
  - "Maria no daba una bofetada a la bruja verde"
- Information extraction (Watson)
- Automatic essay grading
- Spell checking, grammar checking

Why is NLP hard?

- We do it all the time, practically without thinking about it!
- Garbled input
  - Noisy waveforms input to speech recognition
  - Distorted images for OCR
  - "Cascaded" errors
    - Cascades in NLP
- **Ambiguity**

At the word level

- Homophones
  - “It’s hard to wreck a nice beach”
- Part of speech
  - Duck! [VB Duck]!
  - Duck is delicious for dinner. [NN Duck] is delicious for dinner.
- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank of the river to fish.
  - I went to the bank of windows and chose the one for “complaints”.

What’s a word?

- Break up by spaces, right?

  Ebay | Sells | Most | of | Skype | to | Private | Investors
  Swine | flu | isn’t | something | to | be | feared

- What about these?

  达赖喇嘛在高雄为灾民祈福
  ليبيا تحيي ذكرى وصول القذافي إلى السلطة
  百貨店、8月も不振　大手5社の売り上げ8～11%減
  टाटा ने कहा, घाटा पूरा करो

- (What's a sentence...?)

At the syntactic level

- PP Attachment ambiguity
  - I saw the man on the hill with the telescope
- Structural ambiguity
  - I cooked her duck.
  - Visiting relatives can be annoying.
  - Time flies like an arrow.

Pragmatics and World Knowledge

- Interpretation of sentences requires context, world knowledge, speaker intention/goals, etc.
- Example 1:
  - Could you turn in your assignments now? (command)
  - Could you finish the assignment? (question, command)
- Example 2:
  - I couldn't decide how to catch the thief. Then I decided to spy on the thief with binoculars.
  - To my surprise, I found out he had them too. Then I knew to just follow the thief with binoculars.
    [the thief [with binoculars]] vs. [the thief] [with binoculars]

Difficult cases…

- Requires world knowledge:
  - The city council denied the demonstrators the permit because they advocated violence
  - The city council denied the demonstrators the permit because they feared violence
- Requires context:
  - John hit the man. He had stolen his bicycle.

Agenda

- Administrivia
- Introduction to Computational Linguistics & applications
- Rule-based & statistical NLP

Application Goals

- Science vs Engineering
  - Understanding the phenomenon of human language
  - Building better applications

- Accurate; minimize errors (false positives/negatives)
- Maximize coverage
- Robust, degrades gracefully
- Fast, scalable

Rule-Based Approaches

- Prevalent through the 80's
  - Rationalism as the dominant approach
- Manually-encoded rules for various aspects of NLP
  - E.g., swallow is a verb of ingestion, taking an animate subject and a physical object that is edible, …

What's the problem?

- Rule engineering is time-consuming and error-prone
  - Natural language is full of exceptions
- Rule engineering requires knowledge
  - Is this a bad thing?
- Rule engineering is expensive
  - Experts cost a lot of money
- Coverage is limited
  - Knowledge often limited to specific domains

More problems…

- Systems became overly complex and difficult to debug
  - Unexpected interaction between rules
- Systems were brittle
  - Often broke on unexpected input (e.g., "The machine swallowed my change." or "She swallowed my story.")
- Systems were uninformed by prevalence of phenomena
  - Why WordNet thinks congress is a donkey…

Problem isn't with rule-based approaches per se, it's with manual knowledge engineering…

The alternative?

- Empirical approach:
  learn by observing language as it's used, "in the wild"
- Many different names:
  - Statistical NLP
  - Data-driven NLP
  - Empirical NLP
  - Corpus linguistics
  - …
- Central tool: statistics
  - Fancy way of saying "counting things"

Advantages

- Generalize patterns as they exist in actual language use
- Little need for knowledge (just count!)
- Systems more robust and adaptable
- Systems degrade more gracefully

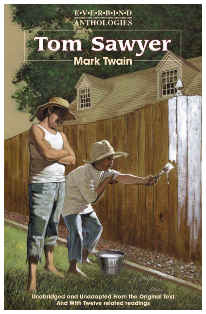
It's all about the corpus!

- Corpus (pl. corpora): a collection of natural language text systematically gathered and organized in some manner
  - Brown Corpus, Wall Street journal, SwitchBoard, …
- Can we learn how language works from corpora?
  - Look for patterns in the corpus

Features of a Corpus

- Size
- Balanced or domain-specific
- Written or spoken
- Raw or annotated
- Free or pay
- Other special characteristics (e.g., bitext)

Grab a "corpus"…



Corpus Characteristics

- Size: ~0.5 MB
- Tokens: 71,370
- Types: 8,018
- Average frequency of a word: # tokens / # types = 8.9
  - But averages lie….

Most Frequent Words (Unigrams)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition

from Manning and Shütze

What else can we do by counting?

Raw Bigram Collocations

Frequency	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York

Most frequent bigram collocations in the New York Times, from Manning and Shütze

Filtered Bigram Collocations

Frequency	Word 1	Word 2	POS
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N

Most frequent bigram collocations in the New York Times filtered by part of speech, from Manning and Shütze

Learning verb "frames"

1	could find a target. The librarian	"showed	off" - running hither and thither w
2	elights in. The young lady teachers	"showed	off" - bending sweetly over pupils
3	ingly. The young gentlemen teachers	"showed	off" with small scoldings and other
4	seeming vexation). The little girls	"showed	off" in various ways, and the littl
5	n various ways, and the little boys	"showed	off" with such diligence that the a
6	t genuwyne?" Tom lifted his lip and	showed	the vacancy. "Well, all right," sai
7	is little finger for a pen. Then he	showed	Huckleberry how to make an H and an
8	ow's face was haggard, and his eyes	showed	the fear that was upon him. When he
9	not overlook the fact that Tom even	showed	a marked aversion to these inquests
10	own. Two or three glimmering lights	showed	where it lay, peacefully sleeping,
11	ird flash turned night into day and	showed	every little grass-blade, separate
12	that grew about their feet. And it	showed	three white, startled faces, too. A
13	he first thing his aunt said to him	showed	him that he had brought his sorrows
14	p from her lethargy of distress and	showed	good interest in the proceedings. S
15	ent a new burst of grief from Becky	showed	Tom that the thing in his mind had
16	shudder quiver all through him. He	showed	Huck the fragment of candle-wick pe

Figure 1.3 Key Word In Context (KWIC) display for the word *showed*.

$$\mathrm{NP}_{agent}\text{ showed off }(\mathrm{PP}[\textit{with/in}]_{manner})$$

$$\mathrm{NP}_{agent}\text{ showed }(\mathrm{NP}_{recipient})\left(\left\{\begin{array}{l}\mathrm{NP}_{content}\\ \mathrm{CP}[\textit{that}]_{content}\\ \mathrm{VP}[\mathrm{inf}]_{content}\\ \textit{how}\ \mathrm{VP}[\mathrm{inf}]_{content}\\ \mathrm{CP}[\textit{where}]_{content}\end{array}\right\}\right)$$

$$\mathrm{NP}_{agent}\text{ showed }\mathrm{NP}[\textit{interest}]\ \mathrm{PP}[\textit{in}]_{content}$$

$$\mathrm{NP}_{agent}\text{ showed }\mathrm{NP}[\textit{aversion}]\ \mathrm{PP}[\textit{to}]_{content}$$

Figure 1.4 Syntactic frames for *showed* in *Tom Sawyer*.

from Manning and Shütze

How is statistical NLP different?

- No need to think of examples, exceptions, etc.
- Generalizations are guided by prevalence of phenomena
- Resulting systems better capture real language use

Three Pillars of Statistical NLP

- Corpora
- Representations
- Models and algorithms

Agenda

- Administrivia
- Introduction to Computational Linguistics & applications
- Rule-based & statistical NLP

HW0:

- Online tonight, due next Thursday before class

Next time:

- Introduction to finite-state models:
  regular expressions, Chomsky hierarchy, automata and transducers