CMSC 723: Computational Linguistics I — Session #1

Introduction to NLP

Jimmy Lin
The iSchool
University of Maryland

Wednesday, September 2, 2009

About Me

Teaching Assistant: Melissa Egan

About You (pre-requisites)

- Must be interested in NLP
- Must have strong computational background
- Must be a competent programmer
- Do not need to have a background in linguistics

Administrivia

- Text:
  - Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, second edition, Daniel Jurafsky and James H. Martin (2008)
- Course webpage:
  - http://www.umiacs.umd.edu/~jimmylin/CMSC723-2009-Fall/
- Class:
  - Wednesdays, 4 to 6:30pm (CSI 2107)
  - Two blocks, 5-10 min break in between

Course Grade

- Exams: 50%
- Class Assignments: 45%
  - Assignment 1 "warm up": 5%
  - Assignments 2-5: 10% each
- Class participation: 5%
  - Showing up for class, demonstrating preparedness, and contributing to class discussions
- Policy for late and incomplete work, etc.

Out-of-Class Support

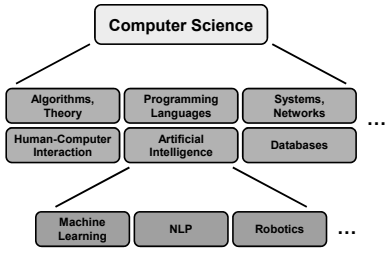
- Office hours: by appointment
- Course mailing list: umd-cmsc723-fall-2009@googlegroups.com

Let's get started!

What is Computational Linguistics?

- Study of computer processing of natural languages
- 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
- Goes by various names
  - Computational linguistics
  - Natural language processing
  - Speech/language/text processing
  - Human language technology/technologies

Where does NLP fit in CS?



Science vs. Engineering

- What is the goal of this endeavor?
  - Understanding the phenomenon of human language
  - Building a better applications
- Goals (usually) in tension
  - Analogy: flight

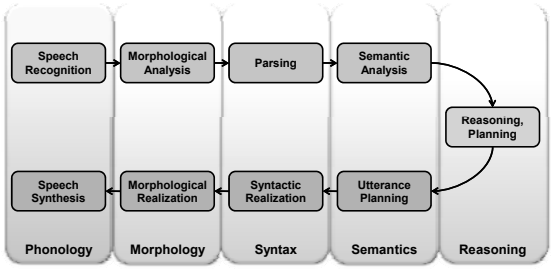
Rationalism vs. Empiricism

- Where does the source of knowledge reside?
- Chomsky's *poverty of stimulus* argument
- It's an endless pendulum?

Success Stories

- "If it works, it's not AI"
- Speech recognition and synthesis
- Information extraction
- Automatic essay grading
- Grammar checking
- Machine translation

NLP "Layers"



Source: Adapted from NLTK book, chapter 1

Speech Recognition

- Conversion from raw waveforms into text
- Involves lots of signal processing
- "It's hard to wreck a nice beach"

Optical Character Recognition

- Conversion from raw pixels into text
- Involves a lot of image processing
- What if the image is distorted, or the original text is in poor condition?

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

Morphological Analysis

- Morpheme = smallest linguistic unit that has meaning
- Inflectional
  - duck + s = $[_{N}$ duck] + $[_{plural}$ s]
  - duck + s = $[_{V}$ duck] + $[_{3rd\ person\ singular}$ s]
- Derivational
  - organize, organization
  - happy, happiness

Complex Morphology

- Turkish is an example of agglutinative language

From the root "uyu-" (sleep), the following can be derived…

uyuyorum	I am sleeping
uyuyorsun	you are sleeping
uyuyor	he/she/it is sleeping
uyuyoruz	we are sleeping
uyuyorsunuz	you are sleeping
uyuyorlar	they are sleeping
uyuduk	we slept
uyudukça	as long as (somebody) sleeps
uyumalıyız	we must sleep
uyumadan	without sleeping
uyuman	your sleeping
uyurken	while (somebody) is sleeping
uyuyunca	when (somebody) sleeps
uyutmak	to cause somebody to sleep
uyutturmak	to cause (somebody) to cause (another) to sleep
uyutturtturmak	to cause (somebody) to cause (some other) to cause (yet another) to sleep
. .	

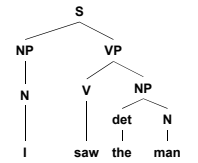
From Hakkani-Tür, Oflazer, Tür (2002)

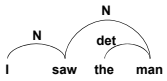
What's a phrase?

- Coherent group of words that serve some function
  - Organized around a central "head"
  - The head specifies the type of phrase
- Examples:
  - Noun phrase (NP): the happy camper
  - Verb phrase (VP): shot the bird
  - Prepositional phrase (PP): on the deck

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:
    $\exists$ x, MAN(x) $\wedge$ SEE(x, I) $\wedge$ TENSE(past)
  - Semantic frames and roles:
    (PREDICATE = see, EXPERIENCER = I, PATIENT = man)

Semantics: More Complexities

- Scoping issues:
  - Everyone on the island speaks two languages.
  - Two languages are spoken by everyone on the island.
- Ultimately, what is meaning?
  - Simply pushing the problem onto different sets of SYMBOLS?

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

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 crook. Then I decided to spy on the crook with binoculars.
  - To my surprise, I found out he had them too. Then I knew to just follow the crook with binoculars.
    [the crook [with binoculars]] vs. [the crook] [with binoculars]

Discourse Analysis

- Discourse: how multiple sentences fit together
- Pronoun reference:
  - The professor told the student to finish the exam. He was pretty aggravated at how long it was taking him to complete it.
- Multiple reference to same entity:
  - George Bush, Clinton
- Inference and other relations between sentences:
  - The bomb exploded in front of the hotel. The fountain was destroyed, but the lobby was largely intact.

Why is NLP hard?

So easy…

Ambiguity

At the word level

- Part of speech
  - [V Duck]!
  - [N Duck] is delicious for dinner.
- Word sense
  - I went to the bank to deposit my check.
  - I went to the bank to look out at the river.
  - I went to the bank of windows and chose the one for “complaints”.

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.

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.

So how do humans cope?

Okay, so how does NLP work?

Goals for Practical 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"
- This approach goes by 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

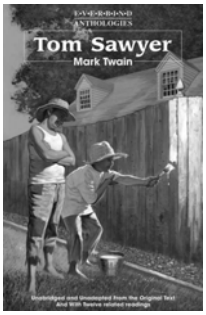
Features of a corpus

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

Getting our hands dirty…

(Example of simple things that you can do with a corpus)

Lets pick up a book…



How many words are there?

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

What are the most frequent words?

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

And the distribution of frequencies?

Word Freq.	Freq. of Freq.
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
50-100	99
> 100	102

from Manning and Shütze

Zipf's Law

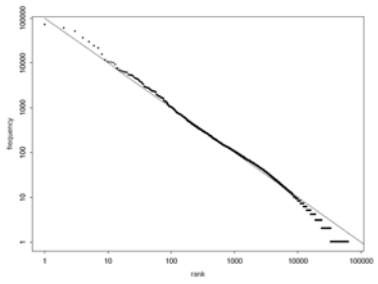
- George Kingsley Zipf (1902-1950) observed the following relation between frequency and rank

$$f \cdot r = c \quad \text{or} \quad f = \frac{c}{r}$$

f = frequency
r = rank
c = constant

- Example: the 50th most common word should occur three times more often than the 150th most common word
- In other words:
  - A few elements occur very frequently
  - Many elements occur very infrequently
- Zipfian distributions are linear in log-log plots

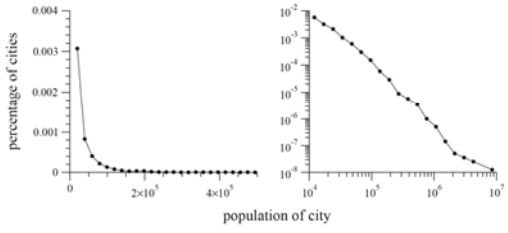
Zipf's Law



Graph illustrating Zipf's Law for the Brown corpus

from Manning and Shütze

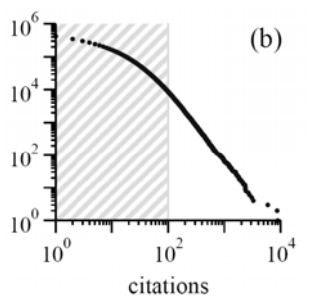
Power Law Distributions: Population



Distribution US cities with population greater than 10,000. Data from 2000 Census.

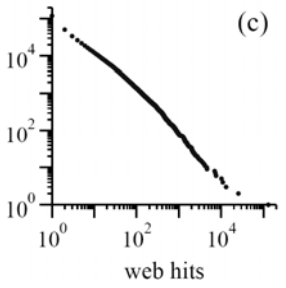
These and following figures from: Newman, M. E. J. (2005) "Power laws, Pareto distributions and Zipf's law." Contemporary Physics 46:323–351.

Power Law Distributions: Citations



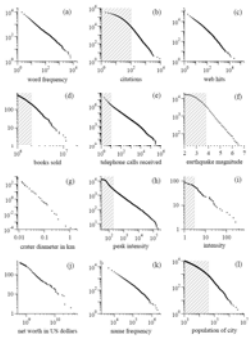
Numbers of citations to scientific papers published in 1981, from time of publication until June 1997

Power Law Distributions: Web Hits



Numbers of hits on web sites by 60,000 users of the AOL, 12/1/1997

More Power Law Distributions!



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 bigrams 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 bigrams 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*.

NP_{agent} showed off (PP[*with/in*]$_{manner}$)

NP_{agent} showed ($NP_{recipient}$) ({ $NP_{content}$ / CP[*that*]$_{content}$ / VP[inf]$_{content}$ / *how* VP[inf]$_{content}$ / CP[*where*]$_{content}$ })

NP_{agent} showed NP[*interest*] PP[*in*]$_{content}$

NP_{agent} showed NP[*aversion*] PP[*to*]$_{content}$

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

from Manning and Shütze

How is this 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

Aye, but there’s the rub…

- What if there’s no corpus available for your application?
- What if the necessary annotations are not present?
- What if your system is applied to text different from the text on which it’s trained?

Key Points

- Different “layers” of NLP: morphology, syntax, semantics
- Ambiguity makes NLP difficult
- Rationalist vs. Empiricist approaches