CS674 Natural Language Processing

- Last class
  - Bayesian method for the pronunciation subproblem in speech recognition
- Today
  - Introduction to generative models of language
    - » What are they?
    - » Why they're important
    - » Issues for counting words
    - » Statistics of natural language

Paradigms in NLP

- Knowledge-based methods
  - Rely on the manual encoding of linguistic knowledge
    - » E.g. FSA's for morphological parsing
- Statistical / learning methods
  - Rely on the automatic acquisition of linguistic knowledge from corpora
    - » E.g. the data-driven, corpus-based methods for automatically deriving linguistic knowledge from text

Statistical/learning-based NLP

- Discriminative models: $P_\Theta(Y \mid X)$
  - X is the input random variable
  - Y is the label random variable
  - Θ is the parameters of the model or model class
- Generative models: $P_\Theta(X) = P(X \mid \Theta)$
  - Bayesian learning procedure
  - Given sample X and model class, find the best set of parameters

Statistical/learning-based NLP

- Ad hoc methods
  - E.g. k-nearest neighbor
  - no clear probabilistic interpretation or justification on the basis of a model

Motivation for generative models

- Word prediction
  - *Once upon a…*
  - *I'd like to make a collect…*
  - *Let's go outside and take a…*
- The need for models of word prediction has not been uncontroversial
  - But it must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term. -Noam Chomsky (1969)
  - Every time I fire a linguist the recognition rate improves. -Fred Jelinek (IBM speech group, 1988)

Why are word prediction models important?

- Augmentative communication systems
  - For the disabled, to predict the next words the user wants to "speak"
- Computer-aided education
  - System that helps kids learn to read (e.g. Mostow et al. critique reading)
- Speech recognition
  - Use preceding context to improve solutions to the subproblem of pronunciation variation

Why are word prediction models important?

- Context-sensitive spelling correction

They are leaving in about fifteen *minuets* to go to her house.
The study was conducted mainly *be* John Black.
The design *an* construction of the system will take more than a year.
Hopefully, all *with* continue smoothly in my absence.
Can they *lave* him my messages?
I need to *notified* the bank of [this problem.]
He is trying to *fine* out.

Why are word prediction models important?

- Closely related to the problem of computing the probability of a sequence of words
  - Can be used to assign a probability to the next word in an incomplete sentence
  - Useful for part-of-speech tagging, word-sense disambiguation, probabilistic parsing

Why are word prediction models important?

- Important in real life situations
- Miss some important words in a conversation, lecture, movie, etc.
- Word prediction gone awry
  - Woody Allen's *Take the Money and Run*
  - Bank teller interprets Woody Allen's sloppily written hold-up note as "I have a gub."

N-gram model

- Uses the previous N-1 words to predict the next one
- In speech recognition, these statistical models of word sequences are referred to as a **language model**

Counting words in corpora

- Ok, so how many words are in this sentence?
- Depends on whether or not we treat punctuation marks as words
  - Important for many NLP tasks
    - » Grammar-checking, spelling error detection, author identification, part-of-speech tagging
- Spoken language corpora
  - Utterances don't usually have punctuation, but they do have other phenomena that we might or might not want to treat as words
    - » I do uh main- mainly business data processing
  - Fragments
  - Filled pauses
    - » *um* and *uh* behave more like words, so most speech recognition systems treat them as such

Counting words in corpora

- Capitalization
  - Should *They* and *they* be treated as the same word?
    - » For most statistical NLP applications, they are
    - » Sometimes capitalization information is maintained as a feature
      - ◆E.g. spelling error correction, part-of-speech tagging
- Inflected forms
  - Should *walks* and *walk* be treated as the same word?
    - » No…for most n-gram based systems
    - » based on the **wordform** (i.e. the inflected form as it appears in the corpus) rather than the **lemma** (i.e. set of lexical l forms that have the same stem)

Counting words in corpora

- **Need to distinguish**
  - **word types**
    - » **the number of distinct words**
  - **word tokens**
    - » **the number of running words**
- **Example**
  - ***All for one and one for all.***
  - **8 tokens (counting punctuation)**
  - **6 types (assuming capitalized and uncapitalized versions of the same token are treated separately)**

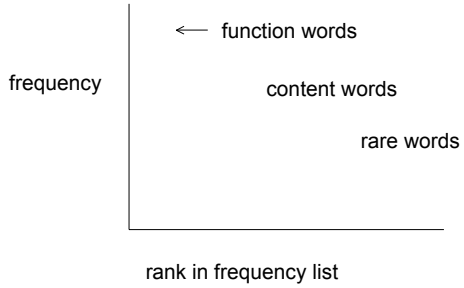
Topics for today

- Today
  - Introduction to generative models of language
    - » What are they?
    - » Why they're important
    - » Issues for counting words
    - » **Statistics of natural language**

How many words are there in English?

- **Option 1: count the word entries in a dictionary**
  - OED: 600,000
  - American Heritages (3rd edition): 200,000
  - Actually counting lemmas not wordforms
- **Option 2: estimate from a corpus**
  - Switchboard (2.4 million wordform tokens): 20,000 wordform types
  - Shakespeare's complete works: 884,647 wordform tokens; 29,066 wordform types
  - Brown corpus (1 million tokens): 61,805 wordform types → 37, 851 lemma types
  - Brown et al. 1992: 583 million wordform tokens, 293,181 wordform types

How are they distributed?



How are they distributed?

- There are stable, language-independent patterns in how people use natural language
  - A few words occur very frequently; most occur rarely
  - In general
    - Top 2 words ~ 10-15% of all tokens
    - Top 6 words ~ 20% of all tokens
    - Top 50 words ~ 50% of all tokens

most common words from *Tom Sawyer*

	word	
1	The	3332
	And	2972
	A	1775
	To	1725
	Of	1440
	…	
	Tom	679
14	With	preposition

Statistical Properties of Text

- The most frequent words in one corpus may be rare words in another corpus
  - Example: "computer" in CACM vs. National Geographic
- Each corpus has a different, fairly small "working vocabulary"

These properties hold in a wide range of languages

Zipf's Law

- Zipf's Law relates a term's frequency to its rank
  - frequency 1/rank
  - There is a constant k such that $freq * rank = k$
  - Rank the terms in a vocabulary by frequency, in descending order

    f_r: frequency of term at rank r

    N : total number of word occurrences

    $p_r = f_r / N$ and
  - Empirical observation $\sum_{r=1}^{V} p_r \approx A \forall r$, $A \approx 0.1$
  - Hence:

$$p_r = \frac{f_r}{N} = \frac{A}{r} \rightarrow rf_r = AN$$

  - $k \approx N/10$ for English

Zipf's Law

Word	Frequency	$r \times p_r$	Word	Frequency	$r \times p_r$
the	1,130,021	0.059	by	118.863	0.081
of	547,311	0.058	as	109,135	0.080
to	516,635	0.082	at	101,779	0.080
a	464,736	0.098	mr	101,679	0.086
in	390,819	0.103	with	101,210	0.091
and	387,703	0.122	from	96,900	0.092
that	204,351	0.075	he	94,585	0.095
for	199,340	0.084	million	93,515	0.098
is	152,483	0.072	year	90,104	0.100
said	148,302	0.078	its	86,774	0.100
it	134,323	0.078	be	85,588	0.104
on	121,173	0.077	was	83,398	0.105

WSJ87 collection (46,449 articles, 19 million term occurrences, 132 MB)

Zipf's Law

- Useful as a rough description of the frequency distribution of words in human languages
- Behavior occurs in a surprising variety of situations
  - English verb polysemy
  - References to scientific papers
  - Web page in-degrees, out-degrees
  - Royalties to pop-music composers
- Zipf postulated a general law regarding human behavior: "principle of least effort"
  - Speaker: small vocabulary is best
  - Hearer: large vocabulary of unambiguous words best
  - Maximally economical compromise solution (Mandelbrot 1954): reciprocal relationship between frequency and rank