LING/C SC/PSYC 438/538

Lecture 23

Sandiway Fong

Administrivia

- Homework 4
  - out today
  - due next Wednesday
  - (recommend you attempt it early)
- Reading
  - Chapter 5: Part of Speech Tagging

Homework 4

- Question 2
  - use corpus file: WSJ9_041.txt
    - a series of *Wall Street Journal* (WSJ) articles from July 26–28 1989
    - approx. 22,000 lines, 150,000 words
  - make sure the text file is properly formatted for your platform:
    - i.e. whether line breaks occur at the right places,
    - e.g. ^M, CR LF etc.

should look something like this…

```
<DOC>
<DOCNO> WSJ890728-0074 </DOCNO>
<DD> = 890728 </DD>
<AN> 890728-0074. </AN>
<HL> Recent SEC Filings </HL>
<DD> 07/28/89 </DD>
<SO> WALL STREET JOURNAL (J) </SO>
<CO> BOW MHC CBK OCTL </CO>
<IN> BOND MARKET NEWS (BON)
STOCK MARKET, OFFERINGS (STK)
FINANCIAL, ACCOUNTING, LEASING (FIN)
INITIAL STOCK OFFERINGS (INI) </IN>
<DATELINE> WASHINGTON  </DATELINE>
<TEXT>
   The following issues were recently filed with the Securities and Exchange Commission:

   Bowater Inc., offering of $300 million of debentures, via First Boston Corp.

   CIT Group Holdings Inc., a unit of Manufacturers Hanover Corp., shelf offering of up to $1.6 billion of
debt securities on terms to be set at the time of sale.
When combined with securities remaining unsold from a previous offering, the filing gives the company as
much as $2 billion of securities available for sale.

   Continental Bank Corp., shelf registration of up to $285 million of preferred stock, via: Shearson
Lehman Hutton Inc.; Goldman, Sachs & Co.; Merrill Lynch Capital Markets; and Dean Witter Reynolds Inc.

   DataImage Inc., initial offering of up to 550,000 common shares, via Coburn &amp;amp; Meredith Inc.

   Octel Communications Corp., proposed offering of 1.5 million common shares, via Alex.
Brown &amp;amp; Sons Inc. and Hambrecht &amp;amp; Quist Inc.

</TEXT>
</DOC>
```

Homework 4

- Part 1 (10 points)
  - **Submit your code** (not the corpus)
  - Use only the text between <text> and </text> markers
    (*write Perl code*)
    - don't include headers etc.
  - Add words <s> and </s> to mark the start and end of each sentence
    (*write Perl code*)

Example:

<s> Sun Microsystems Inc. said it will post a larger-than-expected fourth-quarter loss of as much as $26 million and may show a loss in the current first quarter, raising further troubling questions about the once high-flying computer workstation maker. </s>

- Use any n-gram package from CPAN (*or roll your own*) to compute unigram and bigram frequencies for this corpus
  - e.g. `count.pl` from Text-NSP
  - ***document what you did***
  - if you use Text-NSP define your tokens (i.e. *what counts as a word*)
  - **Note**: in Text-NSP hyphens are non-word separators and deleted, e.g. Bristol-Myers = two words

Homework 4

- Part 2 (10 points)
  - What are the most frequent and 2nd most frequent proper nouns in the corpus?

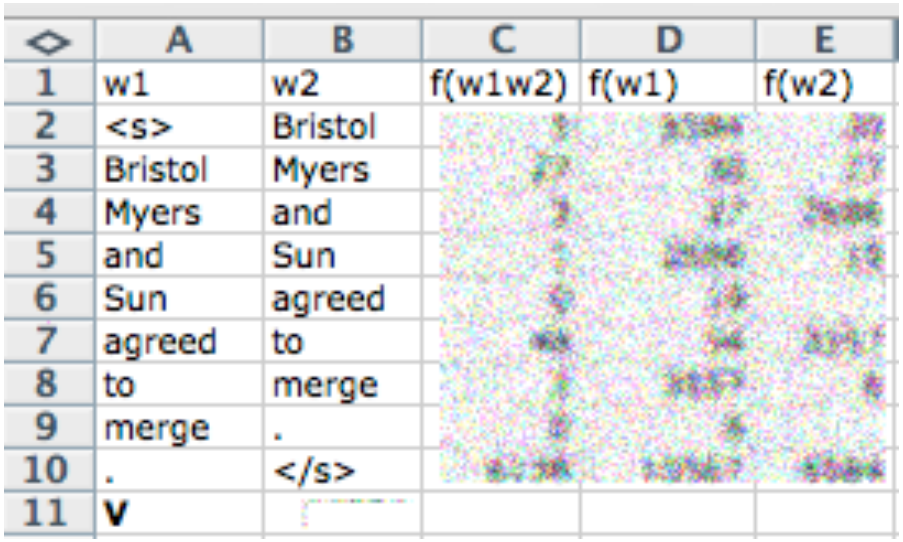
    (*define what you mean by the term "proper noun"*)

  - What are the most frequent auxiliary and non-auxiliary* verb forms?

    *(Take auxiliary verb to mean forms of auxiliary *be*, modals, *do*)

Homework 4

- Part 3 (20 points)
  - compute the probability of the two sentences:
    1. `<s> Bristol-Myers agreed to merge with Sun . </s>`
    2. `<s> Bristol-Myers and Sun agreed to merge . </s>`
  - **which one would you expect to have higher probability?**
  - use the bigram approximation
  - use add-one smoothing to deal with bigram zeros
  - show how you computed your answer: use tables populated with the relevant bigram frequencies and probabilities to document your work
  - submit your excel spreadsheet

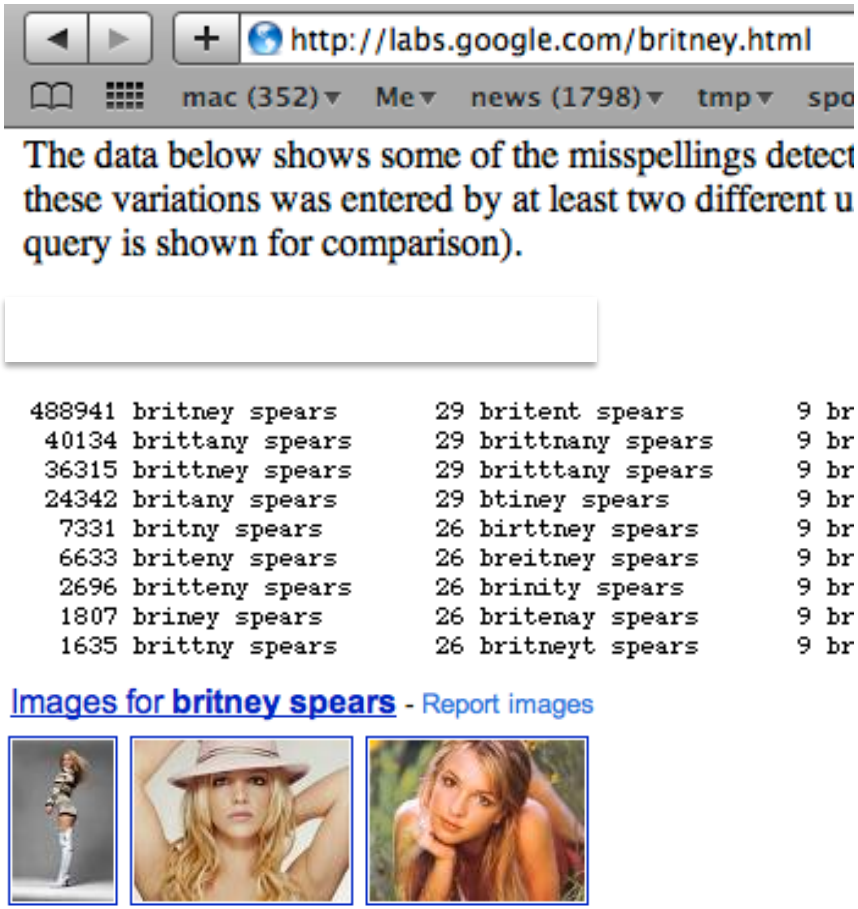


	A	B	C	D	E
1	w1	w2	f(w1w2)	f(w1)	f(w2)
2	<s>	Bristol			
3	Bristol	Myers			
4	Myers	and			
5	and	Sun			
6	Sun	agreed			
7	agreed	to			
8	to	merge			
9	merge	.			
10	.	</s>			
11	V				

Homework 4

- Part 4 (10 points)
  - What are the minimum edit distances for the top 5 misspellings of *Britney*?
  - assuming
    1. unit cost for add, delete and substitute operations
    2. cost 1 for add and delete, cost 2 for substitute
- Part 5 (15 points)
  - how would you modify edit distance to obtain the correct ordering for the 5 misspellings?
  - e.g. cost(*brittany*) < cost(*brittney*) < cost(*britany*) etc.

- in both parts, show your work: submit your spreadsheet(s)



Last Time

Maximum Likelihood Estimation
Relative frequency
$p(w_n|w_{n-1}) = f(w_{n-1}w_n)/f(w_{n-1})$

Sample space: w_{n-1}…
Event: $w_{n-1}w_n$

Language Models and N-grams

- **Given a word sequence**
  - $w_1\, w_2\, w_3 \ldots w_n$
- **chain rule**
  - *how to compute the probability of a sequence of words*
  - $p(w_1\, w_2\, w_3 \ldots w_n) = p(w_1)\, p(w_2|w_1)\, p(w_3|w_1w_2) \ldots p(w_n|w_1 \ldots w_{n-2}\, w_{n-1})$
- **Bigram approximation**
  - *just look at the previous word only* (*not all the proceedings words*)
  - **Markov Assumption**: **finite length history (**1st order Markov Model)
  - $\mathbf{p(w_1\, w_2\, w_3 \ldots w_n) \approx p(w_1)\, p(w_2|w_1)\, p(w_3|w_2) \ldots p(w_n|w_{n-1})}$
- **Trigram approximation**
  - 2nd order Markov Model: *just look at the preceding two words only*
  - $\mathbf{p(w_1\, w_2\, w_3 \ldots w_n) \approx p(w_1)\, p(w_2|w_1)\, p(w_3|w_1w_2) p(w_4|w_2w_3) \ldots p(w_n|w_{n-2}\, w_{n-1})}$

Colorless green ideas

- **examples**
  - (1) **colorless green ideas** sleep furiously
  - (2) furiously sleep ideas green colorless
- **Chomsky** (1957):
  - . . . It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, **in any statistical model for grammaticalness**, these sentences will be ruled out on identical grounds as equally `remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.
- **idea**
  - (1) is syntactically valid, (2) is word salad
- Statistical Experiment (Pereira 2002)

Colorless green ideas

- **examples**
  - (1) **colorless green ideas** sleep furiously
  - (2) furiously sleep ideas green colorless
- Statistical Experiment (Pereira 2002)

$$p(w_1 \cdots w_n) = p(w_1) \prod_{i=2}^{n} p(w_i|w_{i-1}) \quad .$$

bigram language model

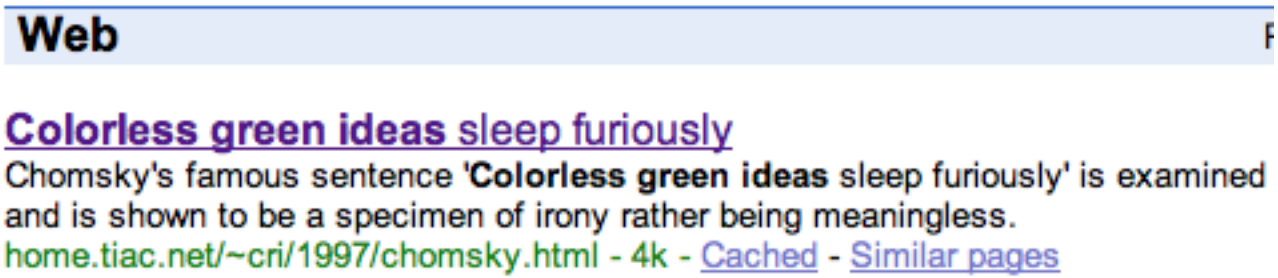
Using this estimate for the probability of a string and an aggregate model with $C = 16$ trained on newspaper text using the expectation-maximization (EM) method (Dempster, Laird, & Rubin, 1977), we find that

$$\frac{p(\text{Colorless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5 \quad .$$

Thus, a suitably constrained statistical model, even a very simple one, can meet Chomsky's particular challenge.

Interesting things to Google

- **example**
  - **colorless green ideas** sleep furiously
- **First hit**



Interesting things to Google

- **example**
  - **colorless green ideas** sleep furiously
- **first hit**
  - **compositional semantics**
  - a **green idea** is, according to well established usage of the word "green" is one that is an idea that is **new and untried**.
  - again, a **colorless idea** is one **without vividness, dull and unexciting**.
  - so it follows that a **colorless green idea** is a **new, untried idea that is without vividness, dull and unexciting**.
  - to **sleep** is, among other things, is to be in a state of dormancy or inactivity, or in a state of unconsciousness.
  - to **sleep furiously** may seem a puzzling turn of phrase but one reflects that **the mind in sleep often indeed moves furiously with ideas and images flickering in and out**.

Interesting things to Google

- **example**
  - **colorless green ideas** sleep furiously
- **another hit**: (*a story*)
  - "So this is our ranking system," said Chomsky. "As you can see, the highest rank is yellow."
  - "And the **new ideas**?"
  - "The **green ones**? Oh, **the green ones don't get a color until they've had some seasoning**. These ones, anyway, are still too angry. **Even when they're asleep, they're furious**. We've had to kick them out of the dormitories - they're just unmanageable."
  - "So where are they?"
  - "Look," said Chomsky, and pointed out of the window. There below, on the lawn, the colorless green ideas slept, furiously.

More on N-grams

- How to degrade gracefully when we don't have evidence
  - Backoff
  - Deleted Interpolation

- N-grams and Spelling Correction

Backoff

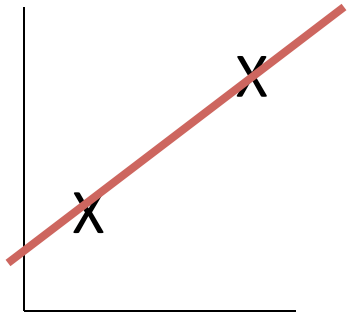
- **idea**
  - Hierarchy of approximations
  - trigram > bigram > unigram
  - *degrade gracefully*
- Given a word sequence fragment:
  - $\ldots w_{n-2}\, w_{n-1}\, w_n \ldots$
- **preference rule**
  1. $p(w_n \mid w_{n-2}\, w_{n-1})$ if $f(w_{n-2} w_{n-1}\, w_n) \neq 0$
  2. $\alpha_1 p(w_n \mid w_{n-1})$ if $f(w_{n-1}\, w_n) \neq 0$
  3. $\alpha_2 p(w_n)$
- notes:
  - α_1 and α_2 are fudge factors to ensure that probabilities still sum to 1

Backoff

- **preference rule**
  1. $p(w_n \mid w_{n-2}\, w_{n-1})$ if $f(w_{n-2} w_{n-1}\, w_n) \neq 0$
  2. $\alpha_1 p(w_n \mid w_{n-1})$ if $f(w_{n-1}\, w_n) \neq 0$
  3. $\alpha_2 p(w_n)$
- **problem**
  - if $f(w_{n-2} w_{n-1}\, w_n) = 0$, we use one of the estimates from (2) or (3)
  - assume the backoff value is non-zero
  - then we are introducing non-zero probability for $p(w_n \mid w_{n-2}\, w_{n-1})$ – which is zero in the corpus
  - then this adds "probability mass" to $p(w_n \mid w_{n-2}\, w_{n-1})$ which is not in the original system
  - *therefore, we have to be careful to juggle the probabilities to still sum to 1*

Deleted Interpolation

- *fundamental idea of interpolation*
- **equation**: (*trigram*)
  - $\mathbf{p}(w_n \mid w_{n-2}\, w_{n-1}) =$
    - $\lambda_1 p(w_n \mid w_{n-2}\, w_{n-1}) +$
    - $\lambda_2 p(w_n \mid w_{n-2}) +$
    - $\lambda_3 p(w_n)$



- **Note**:
  - λ_1, λ_2 and λ_3 are fudge factors to ensure that probabilities still sum to 1

Part of Speech (POS) Tagging

JM Chapter 5

- Parts of speech
  - Traditional 8:
    - e.g. englishclub.com =>
  - traced back to Latin scholars, back further to ancient Greek (Thrax)
  - not everyone agrees on what they are ..
    - Textbook lists:
    - open class 4 (noun, verbs, adjectives, adverbs)
    - closed class 7 (prepositions, determiners, pronouns, conjunctions, auxiliary verbs, particles, numerals)
  - or what the subclasses are
    - e.g. *Proper Noun in the last homework*
    - Textbook answer below

part of speech	function or "job"	example words	example sentences
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats **quickly**. When he is **very** hungry, he eats **really** quickly.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	joins clauses or sentences or words	and, but, when	I like dogs **and** I like cats. I like cats **and** dogs. I like dogs **but** I don't like cats.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

* Some grammar sources categorize English into **9** or **10** parts of speech. At EnglishClub.com, we use the traditional categorization of **8** parts of speech. Examples of other categorizations are:

- Verbs may be treated as two different parts of speech:
  - **Lexical Verbs** (*work, like, run*)
  - **Auxiliary Verbs** (*be, have, must*)
- **Determiners** may be treated as a separate part of speech, instead of being categorized under Adjectives

Proper noun
Common noun

Nouns are traditionally grouped into **proper nouns** and **common nouns**. Proper nouns, like *Regina*, *Colorado*, and *IBM*, are names of specific persons or entities. In English, they generally aren't preceded by articles (e.g., *the book is upstairs*, but *Regina is upstairs*). In written English, proper nouns are usually capitalized.

Part of Speech (POS) Tagging

- Uses
  - POS information about a word
    - Pronunciation: e.g. *are you conTENT with the CONtent of the slide*?
    - Possible morphological endings: e.g. V+*s*/*ed*(1)/*ed*(2)/*ing*
  - distributional information about sequences of POS-tagged words
    - e.g. DT [NN/*VB/JJ]
    - (shallow or 1st stage) of parsing
    - Word sense disambiguation (WSD) task: e.g. *bank*

In computational linguistics, the Penn Treebank tagset is the most commonly used **tagset** (*reprinted inside the front cover of your textbook*)

Penn Treebank Part-of-Speech Tags

Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential "there"	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, preterite (past tense)	*ate*
IN	preposition or subordinating conjunction	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama, snow*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	$
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

45 tags listed in textbook

36 POS + 10 punctuation

POS Tagging

- **Task**:
  - assign the right part-of-speech tag, e.g. noun, verb, conjunction, to a word in context
  - **in NLP**: assign one of the 45(48) Penn tags

- **POS taggers**
  - need to be *fast* in order to process large corpora
    - *time taken should be no more than proportional to the size of the corpora*
  - POS taggers try to assign the correct tag without actually (fully) parsing the sentence
    - the *walk* : **noun** I took …
    - I *walk* : **verb** 2 miles every day

How Hard is Tagging?

- Easy task to do well on:
  - naïve algorithm
    - assign tag by (unigram) frequency
  - 90% accuracy (Charniak *et al*., 1993)

- Brown Corpus (Francis & Kucera, 1982):
  - 1 million words
  - 39K distinct words
  - 35K words with only 1 tag
  - **4K with multiple tags** (DeRose, 1988)

That's 89.7%
from just considering
single tag words,
even without getting
any multiple tag
words right

Penn TreeBank Tagset

- standard tagset (for English)
  - *48-tag subset of the Brown Corpus tagset (87 tags)*
  - http://www.ldc.upenn.edu/Catalog/docs/LDC95T7/cl93.html
- Simplifications
  - Tag `TO`:
    - infinitival marker, preposition
    - I want *to* win
    - I went *to* the store

Table 2:
The Penn Treebank POS tagset

1. CC Coordinating conjunction	25.TO to
2. CD Cardinal number	26.UH Interjection
3. DT Determiner	27.VB Verb, base form
4. EX Existential there	28.VBD Verb, past tense
5. FW Foreign word	29.VBG Verb, gerund/present participle
6. IN Preposition/subord.	30.VBN Verb, past participle
218z conjunction	
7. JJ Adjective	31.VBP Verb, non-3rd ps. sing. present
8. JJR Adjective, comparative	32.VBZ Verb, 3rd ps. sing. present
9. JJS Adjective, superlative	33.WDT wh-determiner
10.LS List item marker	34.WP wh-pronoun
11.MD Modal	35.WP Possessive wh-pronoun
12.NN Noun, singular or mass	36.WRB wh-adverb
13.NNS Noun, plural	37. # Pound sign
14.NNP Proper noun, singular	38. $ Dollar sign
15.NNPS Proper noun, plural	39. . Sentence-final punctuation
16.PDT Predeterminer	40. , Comma
17.POS Possessive ending	41. : Colon, semi-colon
18.PRP Personal pronoun	42. (Left bracket character
19.PP Possessive pronoun	43.) Right bracket character
20.RB Adverb	44. " Straight double quote
21.RBR Adverb, comparative	45. ` Left open single quote
22.RBS Adverb, superlative	46. " Left open double quote
23.RP Particle	47. ' Right close single quote
24.SYM Symbol	48. " Right close double quote
(mathematical or scientific)	

48 tags listed here
36 POS + 12 punctuation